TIGHT BINDING BOOK

UNIVERSAL
LIBRARY

OU_160109

UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 530.1 Accession No. G10695

Author P37L

Title Peierls, R.E
Laws of Nature.

This book should be returned on or before the date last marked below.

The Laws of Nature

The Laws of Nature

R. E. PEIERLS
C.B.E., F.R.S.
Professor of Mathematical Physics,
University of Birmingham

NEW YORK
CHARLES SCRIBNER'S SONS
1956

First printed in Great Britain
This edition printed in the United States

To
The Treasurer of Somerville College, Oxford,
and
The Bursar of Gonville and Caius College, Cambridge,
but for whom this book would
not have been written.

PREFACE

THE plan of this book developed when, after the second world war, I was frequently lecturing to various audiences on atomic energy. I was then struck by the fact that, besides the obvious questions about the nature and dangers of atomic weapons, there were always some people with questions about fundamental physics. How did the uncertainty principle relate to this or that? What held an atomic nucleus together? What was a meson, and how could it be related to nuclear forces if there was not enough energy in the nucleus to produce a meson?

Evidently some members of these audiences were infected with the scientific curiosity which makes scientists study the laws of nature, and I was eager to answer their questions. Scientists are employed by the community to try to solve problems about the laws of nature. If the answers are in demand it is their job to provide them. At the same time, it is always a pleasure to be able to pass on information to people who are anxious to know.

However, one could evidently not get very far on these subjects in the few minutes available for answering questions. A question might depend on some part of quantum theory, and one could hardly expect the questioner to be familiar with this discipline. Even had there been time to explain the principles of quantum theory, one would have had to rely on some knowledge about atoms, or about waves. The answer to one question would soon have developed into a course of lectures. At the same time, I also found that wherever a question could be answered without going too far back to the beginning, the need to keep the language simple and avoid mathematics was no obstacle. In fact, I often found that the experience of stating an argument in simple, non-technical terms helped me to understand it more clearly myself than when I had looked at it only in its mathematical form.

This led to the idea of trying to set out the principles of modern physics in simple language and without assuming any previous knowledge. My experience led me to believe that it was possible to do this in a reasonable space, and that in doing so I would clarify my own ideas. Of these predictions I have already found the second one correct. Whether I was right in the first I must leave the reader to judge.

The book aims to be intelligible to the non-specialist. No particular knowledge of science is assumed, but those who possess some knowledge of physics will find it easier to follow. This applies perhaps more strongly to the first two chapters. Since mechanics and electricity are familiar to many readers, I have set out the arguments in those subjects in greater detail, thereby giving examples of the kind of reasoning which, later on, has to be treated more superficially. It may therefore be that some readers who are less accustomed to the reasoning of the physicist will find parts of the first two chapters a little involved. My advice to them would be to read on and to return, if necessary, for any points quoted in the later chapters that appear to them essential.

There may also be readers who know more mathematics than I am using; but to introduce a little more mathematics would, in fact, hardly help, unless one could assume a thorough understanding of differential equations. For anyone whose mathematics reaches that far, more technical books would be preferable. But I have found also that it does no harm even for the student of physics to see the arguments presented in a simple descriptive way. I have, in fact, in advanced lectures to students, used with advantage some of the illustrations worked out in the course of writing.

I would regard the effort of writing as well rewarded if I had succeeded in showing to some readers who have never been introduced to science at all, a little of the aims, the methods and the conclusions of the scientist. In these days of specialised education, one hears a good deal about the narrow-minded scientist or technician whose education has neglected all human values, together with the arts and humanities. It may well be that he exists, though he would be quite unlike most of the scientists and technicians of my acquaintance. But I am certain that there exists also the other extreme, the person whose education has concentrated on the arts and humanities and whose mind is closed to any interest in science. In fact, there are still many educators who regard science as having little educational value. Their ideas of broadening the basis of education for the scholar of humanities tend to include subjects like the history of science, or the philosophy of science, as more respectable than the teaching of science as such. These are fascinating subjects, but I do not believe that it is profitable to study them except on the basis of an understanding of the principles of science itself. Without this, they resemble an attempt to teach the history of art to a person who has never seen a picture, or the theory of music to the deaf. I believe, therefore, that in any attempt to introduce some aspect of science into the education of the non-scientist, the kind of approach I have chosen in this book should be suitable.

With the completed typescript in front of me, I am strongly aware of all that has been omitted. Apart from the limitations in the subject matter, space has not made it possible to give much detail of the tools of the physicist and of the practical problems of what we must always remember to be an experimental science. I hope I have not allowed the reader to forget the way in which all the conclusions at which we have arrived depend on empirical knowledge and on experimentation. But the book conveys no picture of the actual life of the experimenter and what goes on in his laboratory. Some of my colleagues, for whom the greatest interest in physics lies in the development and use of ingenious new equipment may, indeed, feel that my emphasis is grossly misplaced.

It was also necessary to picture the development of ideas as more direct than it has been in history. I could only describe the ideas which proved themselves on test and not the far more numerous ones which were disproved or proved unconstructive and had to be abandoned. Equally, I could not attempt to convey any feeling of the experience of watching these new concepts grow and prove themselves, an experience well illustrated in the recent Reith lectures of Dr. J. R. Oppenheimer, who, in turn, was blamed by many critics for not doing what this book attempts to do.

Finally I should acknowledge many debts. The thoughts presented in the book are in no way original or personal but represent the normal equipment of a modern physicist. I should not present them without remembering many teachers including Heisenberg, Pauli and above all Niels Bohr, as well as numerous colleagues, who have helped me to become acquainted with the subject and achieve some degree of clarity. Many friends, young and old, have helped me by acting as guinea-pigs when I tried parts of the book on them and they all have made valuable comments. I should like to acknowledge also much constructive criticism of the text from Professor Lancelot Hogben.

CONTENTS

page

PREFACE 7

INTRODUCTION 15

1 MOTION AND FORCE

Newton's Laws 19
Gravity and Other Forces 23
Momentum and Energy 25
Collisions 31
Angular Momentum 34
Starting Conditions and Degrees of Freedom 38

2 ELECTRICITY AND MAGNETISM

Coulomb's Law 41
Lines of Force 45
Field Laws or Action at a Distance 50
Electromagnetism 52
Induction. Generator and Transformer 57
Completion of Laws 61
Electromagnetic Waves 64
Field Energy 68

3 LIGHT

Spectrum. Interference. Waves 70
Diffraction, Polarization, Speed 75
Light Consists of Electromagnetic Waves 78
Emission Theory. Geometrical Optics 80

4 ATOMS AND ELECTRONS

Chemistry and the Atomic Hypothesis 87
The Size of an Atom. Ions 90
The Electron 93
What are Atoms made of? 95
Why Don't Atoms Collapse? 98
Many Facts Can be Understood 99

5 ATOMS IN CROWDS: HEAT AS IRREGULAR MOTION

Quantity of Heat. Heat and Disorder 102
Heat and Atomic Motion 105
Certainty and the Laws of Chance 109

page
More Tests and Some Difficulties 112
Solids and Liquids 114
Radiant Heat 118

6 RELATIVITY
General Remarks 121
Motion and Rest 123
The Lorentz Contraction 125
How to Compare Lengths and Times 128
Composition of Velocities 133
Mechanics of Fast-Moving Objects 134
Applications and Confirmation 138
Principle of Equivalence 141
General Relativity 142

7 QUANTA, WAVES AND PARTICLES
The Quantum Hypothesis. Photons 146
Electron Diffraction. Standing Waves 152
Variable Wavelength 157
The Hydrogen Atom 161
Particles and Waves 164
The Uncertainty Principle 169

8 THE BEHAVIOUR OF ATOMS
Hydrogen and Helium 177
Other Light Atoms. The Exclusion Principle 180
Angular Momentum 183
The Atomic Beam Experiment. Spin 189
Electron Shells, Ionic Molecules 193
Other Types of Molecules 196
Solids, Metals, Liquids 199
Passage Through a Potential Barrier 205

9 ELECTRONS AT HIGH SPEED
Negative Energies. A Difficulty 210
Spin. Positrons 213
Forces at High Speed. More Difficulties 218
New Methods. Lamb Shift 220

10 THE ATOMIC NUCLEUS
Radioactivity. Alpha-Decay 225
Wave Mechanics Solves Another Paradox 229
Splitting the Nucleus 231
New Equipment 233

page
Neutrons. Isotopes. Mass Defect 237
Nuclear Forces 240
Shell Model. Excited States 244
Beta Rays 245
Fission. Sources of Energy 249

11 MESONS AND OTHER NEW PARTICLES

Cosmic Rays and Tools for their Study 255
Known Particles 259
Mesons 261
Yukawa. More Mesons 264
Beyond the Mesons. A New Chapter Opens 269

WHERE DO WE STAND? 273

INDEX 279

INTRODUCTION

Of all branches of science the study of the structure of matter and of the laws which determine natural phenomena has always had a particular fascination. After all, the spirit of scientific enquiry has its origin in the curiosity of a little boy who must take a watch to pieces to see what makes it work, and the search for the basic laws is just this attitude carried to its logical conclusion.

Having mastered the mechanism of the watch we go on to ask what makes the cog wheels, or the springs, or the luminous numbers on the dial, behave as they do, and this brings us directly into problems concerned with the structure of matter. Originally such problems were the concern both of the physicist and the chemist, but since modern atomic theory has shown that the laws of chemistry are consequences of the laws of physics, the further advance towards the basic laws is left to the physicists.

When we speak here of matter we mean by this only inanimate matter. The laws of nature which will be discussed throughout this book do not include a description of life or of living beings. We are not concerned with the question whether it is likely, or even conceivable, that in future the behaviour of living matter may be described by laws which are as complete and as quantitative as those which now describe the behaviour of watches or of atoms.

In this book I shall try to sketch what we have learnt about the structure of matter and about the underlying physical laws, and I shall try to demonstrate the kind of evidence which makes us feel sure that our ideas are on the right lines.

One danger in any brief outline of a branch of science deserves mention. In the interests of simplicity it is often necessary, and it will be necessary in this book, to quote only one, or only a few facts by way of proof for a new and perhaps surprising theoretical idea. The story therefore acquires something of the nature of a detective story, in which just the one important fact establishes the identity of the murderer, and which without this would present an insoluble problem.

In fact the writer of a detective story invites his reader to construct an alternative solution that would fit all the clues which are given, and in the same spirit readers of popular books about science occasionally invent some theory which would seem to account for the key pieces of evidence as accurately as the current physicists' view, or maybe even better. This is particularly common with developments like that of relativity or of the quantum theory, which seem to conflict with common sense. It is only natural that the reader should try (as all physicists tried) to find a less revolutionary theory that would fit the data.

But the data consist of much more than the few typical examples to which a brief account such as this must be limited. An argument in physics is never accepted as valid on the strength of a single experimental fact. A single experiment often is of vital importance in drawing attention to the need for revising our views. It may be sufficient to suggest a new explanation, but this new explanation will still be regarded with an open mind until further facts come along which fit in with it, and until further predictions based on the new hypothesis are confirmed in so many ways that we are convinced of its reliability.

Anyone who wants to question the methods of science must therefore acquaint himself not merely with a few examples, but with enough of the evidence in detail to be able to judge the strength of the case.

The reader of this book is in a similar position to one who has read a newspaper account of a trial. This will show him the kind of evidence on which the verdict was based, but if he wants to claim that the jury were mistaken he must go through all the technicalities of the full evidence on which the jury formed their opinion.

At this point the reader perhaps objects: if physicists are really as careful as that in forming their opinions, how is it that their views of the fundamental laws can be subject to such drastic changes as in the case of relativity, which destroyed the results of Newton's mechanics, or in the case of the theory of light quanta replacing the wave theory of light? How could scientists believe for centuries in theories which were later found incorrect?

This is an important question. The answer is that physics (or for that matter any other empirical science) cannot claim that its results have absolute and final truth. What can be claimed is that we have a picture which is in many respects very close to the truth, so close in fact that for most purposes the differences are unimportant. As time goes on the laws of physics which we have formulated are subject to tests of greater accuracy, or are tested in regions which are much wider than those for which the law was originally developed. In this way a need for revision is sometimes discovered, and the law has to be widened so as to include phenomena which were not previously suspected; but its first form remains a valid law for all practical purposes for the phenomena for which it was first derived.

When, in the later chapters, we review the most revolutionary recent changes in our concepts we shall see that in fact the changes do not mean rejecting the older laws as incorrect, but only recognizing them as incomplete.

This process of gradually correcting the laws with which we describe natural phenomena is a very important feature of the nature of science. Perhaps a simple analogue may help to illustrate the point. If we look at the photograph of a landscape in snow, showing part of a snowfield in sunlight and another part in the shade, we would describe this photograph as having one area in a very light shade of grey, almost white, and another uniform area of a slightly darker shade of grey. On a more careful study of the same photograph under a miscroscope we discover that the grey areas are in fact made up of small black dots, the silver grains of the photographic process, on a white background, the darker area differing from the lighter by having more of the black dots.

In a sense this discovery has proved the first description wrong, but it would be more reasonable to say that the new description refines the old one and replaces it when we are concerned with far greater detail than was the case at first. The old description is still good enough when we are concerned with taking a photograph or with looking at it. In fact by being simpler it is more valuable for that purpose. If we had to think of photographs always as collections of black points on a white background we should find the photographs quite useless as pictures.

In the same way Newton's laws of mechanics and the wave theory of light retain their value for most practical problems within our everyday experience, in spite of our knowledge that the truth they contain is not the whole truth.

In the course of this book I shall try to sketch the way in which discoveries of the past fifty years have led us to widen and to correct our ideas about the laws of nature, and to show the stage which has now been reached.

In telling the story I shall not attempt to follow the development in its actual historical order. In many cases, of course, the reader will have to be taken through the same steps of reasoning which also were followed in the actual development of physics, but in other cases it is simpler to bring in new ideas at an earlier stage, or to save time by cutting out detours.

1

Motion and Force

NEWTON'S LAWS

MECHANICS is by far the oldest branch of physics. A good deal of it is taught at school, and probably most readers will be familiar with the laws of mechanics. It is often forgotten that mechanics is a branch of natural science and that its laws come to us from observation. By our experience and by our education we are so familiar with these laws that we tend to take them for granted, and in England in particular mechanics is often regarded as a branch of mathematics.

Nevertheless it will be useful to review its principles briefly in this chapter, both because they are the earliest and simplest example of the application of quantitative study to natural phenomena and also because later on we shall have to discuss the changes in these principles which are necessary to apply them to the motion of very small objects and to motion at very high velocity. It will then be necessary to revise ideas which appear to us at first sight quite obvious and unchangeable, and we must therefore understand the facts on which they are based.

Mechanics is concerned with the motion of bodies and with the forces which cause them to move. When we say a body moves, this is another way of saying that it is changing its position and we therefore meet at once the idea of a rate of change. Thus the velocity or speed of a particle is the rate at which its position changes. In the special case of a body moving uniformly, or with constant velocity, the velocity is the ratio of the distance covered to the time taken to cover it, as in the case of a race, when the speed of a runner, or a horse, or a plane, is usually determined by timing with a stop-watch the transit over a measured distance. However, if the speed is not constant we shall get different answers according to the length of the time interval over which we check the run. In that case to define a speed at some instant we must imagine that by very accurate observation we measure accurately the small distance which the body has covered in a very short interval of time. As the mathematician would say, the speed is the limit for an infinitely small time interval of the ratio of the distance to the time interval.

The definition of such a rate of change is the basic concept of differential calculus, and most of the quantitative laws of physics can best be stated in the form of equations involving differentiation, or as we call it, of differential equations. The reader of this book will not, however, be assumed to be thoroughly at home with calculus, and we shall avoid its use, even though this will often require a clumsy statement in words of an argument which an equation could present much more simply, while some ideas have to be left out as being unintelligible without calculus.

While we can to some extent avoid the formulae of calculus, we cannot do without the idea of a rate of change. However, in this mechanical age the idea of the speed of a body at a given moment is familiar enough. While it may be a strain on the imagination to think of the "measured mile and stop-watch" method carried out with an extremely small time interval, the reader will be familiar enough with watching the changing speed of a motor-car indicated by its speedometer.

The most fundamental law of mechanics is then the inertial law of Galileo, which states that if left to itself a body will move with uniform velocity in one and the same direction.

This law seems nowadays almost obvious, but it is quite contrary to everyday experience in which we observe that bodies in general tend to stop moving when the force that has caused their motion ceases to act. This is, of course, explained by saying that bodies are, as a rule, subject to the forces of friction with other bodies, or the resistance of the air, and if these were removed the bodies would continue to travel.

It is easily shown that if we reduce the causes of friction by making our test objects slide or roll on very smooth surfaces, and if we diminish the air resistance by shaping the bodies suitably, or by evacuating the space in which they move, they will come much nearer to the ideal of Galileo's law. This shows that we are right in regarding the tendency of bodies to stop moving, not as a natural tendency inherent in them, but as due to secondary causes.

Whenever the velocity of a body changes, i.e. when its motion is accelerated, we therefore say this must be due to some force, and the relation between force and acceleration is described by Newton's "second law," (the first being Galileo's principle) which states that the force equals mass times acceleration.

By the acceleration we again mean a rate of change, this time the rate of change of the velocity with time. We can take as our demonstration object again the motor-car. The sensation of a car accelerating rapidly (or braking hard, which means an acceleration in the opposite direction) is familiar to all of us; it is not hard to think of this acceleration as something expressible in quantitative terms.

People sometimes argue whether Newton's second law is a definition of force or of mass, or whether it is a statement of an objective fact. It is really a mixture of all these things; and as this is a very typical situation for a physical law, it is worth making this point quite clear.

Newton's law states, first of all, that the acceleration of a given object by a given force is always the same, independent of how fast the object is moving and also independent of other circumstances, such as geographical position, time of the year, temperature, etc. This is a statement of fact which could be proved or disproved by observation. It further states that if the same force is applied to different objects their accelerations will always be in the same ratio. For example if one object *A* is accelerated twice as much as another object *B* by a certain force, then if we apply any other force it would always give *A* twice the acceleration that it gives to *B*. This again is a statement of fact. In this way we can attribute to each body a quantity which we might call the inertia, and we can determine by observation the ratio of the inertia of any two objects. If we did this we would find that the inertia is additive, i.e. if we combined two objects together the resulting body will have an inertia which is the sum of the inertia of the separate objects. This makes it reasonable to use, in place of inertia, the term mass, which we have thereby defined. It is, however, defined on an arbitrary scale, that is, we must specify some object and call its mass one pound or one kilogramme and then the masses of other objects can be expressed as multiples or as fractions of this unit.

All this refers to some given force, and the careful reader will have noticed that for my argument it is necessary to be able to apply equal forces to different objects. This would be difficult without the use of Newton's third law, which states that action equals reaction, or in other words that if one object *A* exerts a certain force on another object *B* then *B* exerts on *A* a force opposite and equal to the first force. This we may regard as part of the definition of a force, since without it any comparison of forces would be very difficult. Using this law we can then find many occasions on which bodies of very different masses interact with each other and thereby carry out the comparisons which we have discussed.

We also find that different forces are additive in the sense that if two forces act on a given body together in the same direction then the acceleration will be the sum of the accelerations that would have been produced by each separately. For instance, if two horses pull equally hard on a cart they will give it twice the acceleration which one horse would have produced. This again is an observable fact and it helps to show the definition of a force to be reasonable.

So far we have not taken any notice of direction, but in fact velocity, acceleration and force are all directed quantities, or "vectors". When we talk about a force we must specify it in magnitude and in direction, and when we talk of adding two forces, which need not pull in the same direction, we mean the resultant force which is obtained when these two forces push on an object simultaneously.

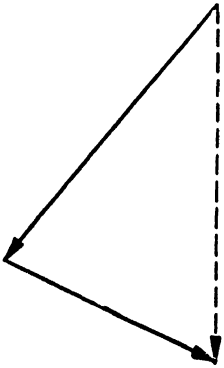


Fig. 1. Composition of forces.

The reader may remember that this resultant force can be constructed by drawing first an arrow in the direction of the first force and of a length corresponding to its magnitude, and then from its end point an arrow in the direction of the second force and of a length corresponding to the magnitude of that. Then the line joining the beginning of the first and the end of the second arrow gives both direction and magnitude of the resultant force. This construction, which is shown in Figure 1, is sometimes known as the triangle of forces. By drawing two more lines to complete a parallelogram, we have the version of this rule referred to as the "parallelogram of forces".

The same principle applies to adding two velocities, and by taking one in the opposite direction we may in the same way take the difference of two velocities.

In this way we must interpret the rate of change of velocity which, as the acceleration, enters Newton's law. A body is therefore also in accelerated motion if its velocity is constant in magnitude, but varies in direction. Thus a force is required to make a particle travel along a curved path with constant speed. To use another motor-car illustration, the passenger in a car going rapidly round a bend is very conscious of the force exerted on him by the seat or the side of the car.

The acceleration of an object moving with constant speed along a curve is $\frac{v^2}{r}$ where v is the velocity and r the radius of curvature. This result can be derived by considering the velocities of the particle at two successive instants, and taking their difference by the triangle or parallelogram rule, but it is rendered plausible by the fact that an acceleration must equal the ratio of a velocity to a time, or in other words must be a distance divided by the square of a time. Now $\frac{v^2}{r}$ is the only combination of the velocity and the radius of curvature which meets this requirement. The direction of this acceleration is at right angles to the direction of motion at the same instant, and it is directed towards the centre of curvature.

To keep a particle moving along the curve one requires therefore a force in the same direction and of magnitude $\frac{mv^2}{r}$ where m is the mass of the object.

Although this force is directed towards the centre, one sometimes refers to a "centrifugal force". This arises because by Newton's third law the force exerted by the object on whatever is enforcing the curvature is in the outward direction. Thus while in a car making a sharp left-hand turn the seat of the car exerts on us a force pushing us to the left, our reaction on the seat is pushing towards the right, and thus away from the centre of curvature.

With this explanation of what is meant by acceleration we have completed our summary of the basic laws of mechanics. For any practical problem, we must of course add to these general laws a specification of the forces which are acting and any limitations which may apply to the motion of the objects under consideration.

GRAVITY AND OTHER FORCES

Of the forces that play a part in mechanical problems of everyday life, one of the most important is the force of gravity which acts on any object vertically downwards. It is proportional to the mass, so that under the influence of gravity alone all objects fall with the same acceleration.

This is another case in which a simple general law of nature seems to contradict common sense. We are intuitively familiar with the fact that a stone will drop more rapidly than a piece of paper, and we cannot recognize the general law until we understand that the cause of the slower falling of paper lies in the air resistance which affects it more than the stone. In other words, the familiar phenomenon of objects falling through air is complicated; the case for which we expect a simple law to be valid is that of objects falling in empty space, to which we can approximate by watching them fall in an evacuated tube.

The fact that the force of gravity on any body, i.e. its weight, is exactly proportional to its mass, or inertia, is at this stage just a surprising coincidence. We shall see later that in the general theory of relativity this fact is given a new meaning linked to the nature of gravitation.

Newton also recognized that the force which makes objects fall towards the earth is only a special case of a general attraction between any two masses, and that this same attraction is responsible for keeping the earth and other planets on their courses around the sun. It is again contrary to our intuition to think that a strong attractive force would not make the earth fall into the sun or the moon into the earth, because we are not accustomed to dealing with bodies subject to so little friction that they can retain their velocity for millions of years. We see in fact that it is possible for an object to revolve in a circular orbit about an attractive centre, provided the attractive force is just of the right magnitude to give it the acceleration which corresponds to the curvature of the orbit. We have just seen that this force is $\frac{mv^2}{r}$. For given radius r and given attractive force there is then always a velocity for which such a steady motion is possible.

The orbits of the planets are not exact circles, and a full description of their motion would require more mathematics than we are going to use, but they are not far from circles and the difference does not alter the nature of the problem. If we compare the period of revolution of different planets and their distances from the sun, observation shows that the square of the period varies as the cube of the radius. Remembering that the velocity v is equal to the circumference of the circle ($2\pi r$) divided by the time of revolution t, we deduce from this observation that the force varies as the inverse square of the distance from the centre.

Reasoning such as this led Newton to formulate the universal law of gravitation, according to which any two bodies of masses m and M, which are a distance r from each other will attract each other with a force $\gamma\frac{mM}{r^2}$ where γ is the constant of gravitation. He showed that for the motion of the planets this law could explain Kepler's rules about the nature of the orbits. The same law, with the same constant, gives correctly the strength of gravity at the surface of the earth.

Later work showed also that small deviations from the exact Kepler orbits could be explained by the mutual attraction of the planets on each other, particularly when they happen to approach closely. All this amounts to a very extensive confirmation of the inverse square law.

The constant γ in this law is not easily determined since we have no direct method of knowing the mass of the sun or any of the planets. The only way of determining γ is to measure directly the attractive force due to some large object whose weight is accurately known. Cavendish did just this by observing the deflection of a sensitive balance on placing some large spheres of lead close to one of the weights, and in this way we know the quantity γ. Knowing this, we may then, from the orbits of the planets, derive the mass of the sun or of the earth.

We have discussed the law of gravitation extensively because it played such a large part in the development of Newton's mechanics, and also because it is the oldest example of a force which obeys a general and simple law.

But there are forces of many other kinds. If this was the beginning of a text-book on mechanics we should give details of the forces which are important in mechanical applications. Examples are the forces between bodies in contact, including contact pressure and friction between them, elastic forces which tend to restore bodies to their normal size and shape, including therefore the force in the spring that drives a watch. Similarly, there are forces due to the pressure of liquids, or gases, including therefore the force of buoyancy that keeps a ship afloat and the pressure of steam in a locomotive, or that of the hot combustion gases in the engine of a motor-car.

But we shall see later that all these types of forces which are connected with the structure of matter are nothing basic in themselves, but can be derived if we know the principles of the structure of matter. Since matter consists of atoms all these forces in the last resort are the consequence of forces between atoms.

Apart from these the only other forces that play a part in everyday experience are electric and magnetic forces, which will be discussed in the next chapter.

MOMENTUM AND ENERGY

In the later discussion it will be convenient to refer to a few other mechanical laws, which are consequences of the laws of Newton. One of these is the law of conservation of momentum. By the momentum of an object we mean its mass times its velocity. The rate of change of momentum is therefore equal to the mass times the rate of change of velocity, which is just the force acting on the object.

Consider then two objects A and B, which exert forces on each other (e.g. they may attract each other). Then the rate of change of the momentum of A is the force exerted by B on A, whereas the rate of change of the momentum of B is the force exerted by A on B. By the law of action and reaction the two are opposite and equal, hence in any given time the momenta of A and B change by opposite amounts, or the sum of their momenta is constant. Here the sum is, of course, to be understood in the sense of the triangle or parallelogram law.

It is easy to see that this result can be extended to any number of bodies, always provided that the only forces present are interactions between those bodies, without any external forces.

We thus recognize the law of conservation of momentum: if in any mechanical system the only forces present are those exerted by some parts of the system on other parts, the momentum of the whole system stays constant.

In the particular case of two small bodies of equal mass the sum of their momenta is $mv_1 + mv_2 = m(v_1 + v_2) = 2m\left(\frac{v_1 + v_2}{2}\right)$. The expression in the bracket is their mean velocity, and this is the same thing as the velocity of a point which is always half-way between the two objects. In this particular case we may therefore state that whatever the forces by which the two objects attract or repel each other, the point half-way between them will always move with uniform velocity in a straight line. If the particles have different masses, this statement is still true of a point which we call the centre of mass. Its position can be calculated in the following way: If the two objects move along a straight line and are, respectively, at distances x_1 and x_2 from a fixed point, then the centre of mass is at a distance $x = \frac{m_1 x_1 + m_2 x_2}{m_1 + m_2}$ from the same reference point. It is easy to see that this centre of mass lies closer to the heavier of the two objects. In the same way one can define the centre of mass of more than two objects, and the total momentum is always equal to Mv_c where M is the total mass of all the objects and v_c is the velocity of their centre of mass.

The next important concept which we shall discuss is that of energy. To understand this let us consider first of all a body falling freely under gravity, assuming its friction in the air again to be negligible. We ask for the relation between the distance through which the object has fallen and the velocity which it has acquired in doing so. It would be a plausible mistake to believe that the velocity should increase uniformly with the distance. This is not correct.

since the force of gravity is constant and by Newton's law the velocity of the object therefore increases uniformly with time. In other words the increase of the speed in each second is the same, but if the speed increases, the distance covered in each second increases, and therefore the same increase in velocity is made over a longer distance. In other words, the velocity increases less than proportionally with the distance. We can make this more quantitative in the following way. Suppose that at time t_1 the object has velocity v_1 and that at a slightly later instant t_2 it has velocity v_2. Then Newtons second law says that $v_2 - v_1 = g\,(t_2 - t_1)$ where g is the constant acceleration of gravity. Now if at the first instant the distance from the starting point is s_1 and at the second instant s_2 then we have $s_2 - s_1 = v\,(t_2 - t_1)$. What is the velocity v which we have to use in this equation? Since the velocity of the particle is not constant, being less in the beginning and more towards the end, it is reasonable to put in the average of the velocity over this time interval, which is $\frac{1}{2}(v_1 + v_2)$. This argument is always valid in an approximate sense for very small time intervals. It happens to be exact for any size of time interval in the case of uniform acceleration, which we had assumed. Expressing therefore in our original equation $t_2 - t_1$ in terms of the distance travelled, we find easily—

$$\tfrac{1}{2}(v_1 + v_2)\,(v_2 - v_1) = g(s_2 - s_1)$$

or multiplying out the first two brackets—

$$\tfrac{1}{2}\,v_2^2 - \tfrac{1}{2}\,v_1^2 = g(s_2 - s_1)$$

In other words, it is not the velocity, but one-half the square of the velocity, which increases by the same amount if the particle has covered the same distance again. Using the fact that the force is mass times acceleration the same result can be written as $\frac{1}{2}\,m(v_2^2 - v_1^2) = F(s_2 - s_1)$ where F is the force. Consider then some particular interval of height, and many objects falling through this interval, which have started from different levels. The ones which have travelled longer and therefore have higher velocities, will increase their velocities by less in passing through this particular interval, but the quantity $\frac{1}{2}mv^2$ will increase by the same amount for each.

From this one can immediately make some useful deductions. Suppose, for example, I throw an object vertically upwards and wait for it to return, with what speed will it reach me? The answer (apart again from air resistance) is: with the same speed as the one with which I threw it, except of course in the opposite direction. This can be seen in this way: I can imagine that I follow its ascent and mark every inch it has risen; as it rises through one particular inch the quantity $\frac{1}{2}mv^2$ will have decreased by a certain amount (namely

F times 1 inch) and as it returns over the same inch the quantity will increase again by the same amount. Hence when it returns to the starting point $\frac{1}{2}mv^2$ has increased as much as it decreased on rising, and must have grown again to what it was at the start.

The quantity $\frac{1}{2}mv^2$ is called the " kinetic energy". Such names for mechanical terms are somewhat arbitrary in origin, and one should not try to attribute to them any meaning related to the non-technical use of the word. For example, the expression "a very energetic person" has probably more analogy with the mechanical concept of power than with that of energy.

We have therefore learnt that the kinetic energy of a freely falling object increases by Fs where F is the force and s the distance the object has fallen. Or if it rises to height h its kinetic energy will decrease by Fh, so that if the particle rises or falls the sum $\frac{1}{2}mv^2 + Fh$ remains constant. This is a special case of the law of conservation of energy. While we refer to the first part as kinetic energy, the second part, Fh, is called "potential energy".

How do these results change when we are not dealing with gravity, but with some other force? First, we note that other forces in general are not the same at every point, but vary in strength. For the moment, let us still restrict ourselves to an object moving in a straight line and let us also assume that the force acting on the object is the same at any given point on this line, no matter when or at what speed the object is passing this point. Then we may use the relation $\frac{1}{2}mv_2^2 - \frac{1}{2}mv_1^2 = F(s_1 - s_2)$ for a very small interval in which the force may be regarded as constant. The right-hand side is then called the "work" done by the force F, and if an object covers any given distance we must add up the total work done by this force by adding together the work done in different intervals. It is therefore again true that when the object returns to the starting point it has passed each interval twice in opposite directions, and the increase in kinetic energy in the one passage is cancelled by the decrease in the other, so that the object returns with its original velocity.

Forces for which this statement is true are called "conservative forces", because for them the law of conservation of energy holds. We can again determine the potential energy by adding up the whole work done by the force as the object moves from the point in question back to a given reference point. Since the only purpose for which potential energy enters into the considerations is to compare the potential energy on two different occasions, it is clear that only differences in potential energy have a unique meaning, and we can count the potential energy from an arbitrary reference point as long as we do so consistently.

However, there are evidently forces which are not conservative, in particular all forces which depend on the velocity of the object. For example friction is a non-conservative force, since it always opposes the direction of motion of the object. If you have an object sliding on a table, it will lose in moving across a certain distance some of its velocity; if later in its motion the object returns over the same distance, it will not regain the kinetic energy it has lost, but, on the contrary, will lose more. It is in fact evident that no velocity-dependent force can be conservative (for an object moving along a straight line), because as the object may well pass the same point in different directions with different velocity, the changes of kinetic energy will not cancel.

Forces which are not conservative may lead to a net increase, as well as a net decrease, of kinetic energy. Of this kind is, for example, the force of the steam on the piston of a steam engine. The steam engine is controlled in such a manner that the steam presses on the piston while it moves out of the cylinder, but as the piston returns the steam is shut off and there is therefore not an equally strong opposing pressure.

When we turn from the restricted case of motion in a straight line to the more general, it is now no longer true that a force dependent only on position is necessarily conservative. This may be seen from the following example. Consider an object exposed to wind blowing over a plain, such that there is no wind on the ground, but that the wind velocity increases with height. Assume also that we have an object on which the wind will exert a force and that this force depends only on the local wind velocity and therefore in the given case only on where the object happens to be. We now attach the object to the end of a string and swing it in a vertical circle, so that at the top of the circle its direction is down-wind. Then in the upper part of the circular motion the object is moving in the same direction as the force of the wind and the work done by the wind is positive, in other words the kinetic energy of the object increases. In the lower half of the circle the object moves in the opposite direction and therefore the wind will oppose its motion, but according to our assumption the wind down below is weaker than above and therefore the loss in kinetic energy due to this will not make up for the increase in the upper half. In ideal conditions therefore the object would gain energy on every round and its speed would increase until it became comparable to that of the wind, when it is no longer reasonable to assume that the wind pressure depends only on the position, but not the velocity of the object.

A force is therefore not conservative even though it depends only on the position of the object, if on describing a complete loop some work is done on the object. This quantity, the work done on describing a certain closed loop, will be called the "circulation". The distinction between conservative and non-conservative forces will be important later on.

Thus, as long as we limit ourselves to purely mechanical considerations the law of conservation of energy is valid only in some very special cases and rarely in any real practical problem, which usually involves engines or living beings as sources of mechanical energy, and friction and other resistances which consume mechanical energy.

However, this situation is profoundly changed if we accept the fact that mechanical energy is only one of many different forms of energy. For example, when the brakes are applied to bring a car to a stop and thereby to destroy its kinetic energy, it is found that heat is generated in the braking system. When the science of heat, which will be reviewed in a later chapter, came to be put on a quantitative basis, so that it was possible to measure a quantity of heat, it was found that the heat produced in destroying a given amount of mechanical energy was always the same. This suggests that we are just exchanging mechanical energy for heat at a certain rate. If we call the heat equivalent of a certain amount of mechanical energy just heat energy, then also friction does not destroy energy, but merely converts mechanical energy into heat energy.

Conversely a steam engine converts heat energy into mechanical energy, but as we shall see later that is not all there is to be said about the problem, since it turns out to be much easier to convert mechanical energy into heat than vice-versa.

There are other forms of energy. One of them is chemical energy. In burning coal, i.e. combining carbon and oxygen into carbon-dioxide, heat energy appears, apparently from nowhere. This comes under the law of conservation of energy if we suppose that matter contains chemical energy according to its chemical state and that the energy of carbon and oxygen separately is higher than that of the product, carbon dioxide. In other words combustion converts chemical energy into heat. In the same way an electric battery converts chemical into electrical, a motor electrical into mechanical and an electric heater, electric energy into heat.

All these forms of energy are capable of being measured quantitatively and for all processes that have been studied the energy always balances. When we come to discuss more recent developments in physics we shall meet many new forms of energy, but we shall find no need to abandon the principle of energy conservation itself, though we shall discuss one particular case in which the question is still open to some doubt. The chances are therefore that the conservation of energy is a universal and valid law of nature, though as with any other empirical law we must be ready to question it if new evidence is discovered which conflicts with it. Only, with a law of such wide validity, any contradictory evidence would have to be very strong and very convincing indeed before it would be accepted.

COLLISIONS

The laws of conservation of energy and momentum are particularly convenient for discussing one type of mechanical process, namely collisions. Since the study of collisions between atomic particles will later be of importance we shall consider now how one uses these conservation laws in simple cases.

First consider an object, let us say a billiard ball, of mass m, moving with velocity v, which collides head-on with another such ball of the same mass, which before the collision was at rest. Now such a collison means that the objects are in contact for a very short time, which depends on the extent to which they will give under pressure, i.e. on their elasticity. Imagine for a moment we were dealing not with billiard balls, but with balloons. If we watch two balloons collide we can see that after they first touch they still go on approaching each other; the parts of their skins which are in contact flatten off and this area of contact grows so that the balloons have a piece of surface in common. The pressure of the gas in either balloon on this surface has therefore more and more effect and this results in a force between them which reduces their velocity towards each other and ultimately forces them apart again.

The situation in the case of billard balls is exactly the same, except that, as they are so much harder, the area of contact will remain very small and the time for which they are in contact will be very short indeed. As a result also the force between them, which must be capable of changing their motion in a very short time, i.e. must produce very high acceleration, is a very strong force indeed. This being so it is usually possible to forget about any other forces such as gravity to which these objects might be subject during this short period of contact, since the changes in velocity which they may produce over such a short time will be small. We can therefore work out their motion during this short interval as if the only acting force was the contact pressure between the two objects.

We have therefore a situation to which the conservation of momentum applies, since the only force is the interaction between the two billiard balls. The momentum immediately before the collison, which is mv, must therefore equal the total momentum after the collision, which is $m(u_1 + u_2)$, if u_1 and u_2 are the velocities of the two balls after the impact. They are in the same direction, since we are considering a head-on collision. In other words, momentum conservation says for this case that the sum of the velocities after the collision equals the velocity of the first ball before. Now the nature of a collison is very different according to whether in this process mechanical energy gets converted into heat or not. This depends on the substances from which the objects are made. Billiard balls approach quite closely to the ideal of perfectly elastic collisions, in which no heat is produced. Then we expect the mechanical energy to be constant and this requires that the kinetic energy just before the collision is equal to that just after. (In between some of the energy is converted into elastic energy and then back again into kinetic energy, but we need not discuss this in detail.) Hence we can write the equation $\frac{1}{2}m(u_1^2 + u_2^2) = \frac{1}{2}mv^2$. Now omitting the common factor $\frac{1}{2}m$ and using the result obtained from momentum conservation this may be written $u_1^2 + u_2^2 = (u_1 + u_2)^2$. Using a well known algebraic rule for the right-hand side we have

$$u_1^2 + u_2^2 = u_1^2 + u_2^2 + 2u_1u_2$$

and this is possible only if u_1u_2 is zero. The product of two numbers can be zero only if at least one of these numbers is zero. So either u_1 or u_2 must be zero and therefore either of the balls must be at rest after the collision. Since u_1 and u_2 must add up to v, the other one must be moving after the collision with velocity v. Now it cannot be the first one which moves in this manner, because for that it would have to pass through the second. We conclude therefore that the first ball must stop and the other go on with velocity v.

It is evident that we have made an essential use of the fact that the masses were equal; we can therefore test whether the masses of two billiard balls are indeed equal by seeing whether the collision brings the first ball just to rest. This is an illustration of the general statement I made earlier on, that Newton's third law, which requires conservation of momentum, allows us to compare the effect of the same force applied to two different objects and therefore directly compare masses without measuring force.

Another extreme case is that of perfectly inelastic bodies in which all the energy that has been put into elastic deformation is converted into heat, and none of it restored to mechanical energy. Now it is clear that the balls will go on piling up elastic deformation only until their relative motion has stopped, i.e. until they are moving with the same velocity. In the case of two equal perfectly inelastic objects of which snowballs or lumps of clay are good examples, we expect that after the collision their velocities will be equal and from momentum conservation equal to half the initial velocity v.

So far we have talked only about "head-on" collisions in which we expect that both balls will move along the straight line along which the first one arrived. Consider now a more general collision as sketched in Fig. 2a. It is clear that the force pushing ball 2 in this case will not project it in the direction which in the drawing is vertical, but will push it more towards the right and that ball 1 will be deflected towards the left, as indicated by the upper arrows. Now momentum conservation still says, as before, that the velocities after the collision must add up to v, but this must now be understood in the sense of adding directed quantities or "vectors", i.e. in the sense of the triangle rule. This is shown in Figure 2b, which shows how we have to picture u_1 and u_2 making up the resultant v. Now if the balls are again perfectly elastic it follows that we must have again $v^2 = u_1^2 + u_2^2$; the triangle of Figure 2b must be such that the square of one side equals the sum of the squares on the two others. This is what the theorem of Pythagoras asserts of a right-angled triangle, and it is true only of a right-angled triangle. Therefore u_1 and u_2 are at right angles to each other. Hence the laws of conservation of energy and momentum show that if a moving object collides perfectly elastically with one at rest, of the same mass, then after the impact the two bodies move at right angles to each other.

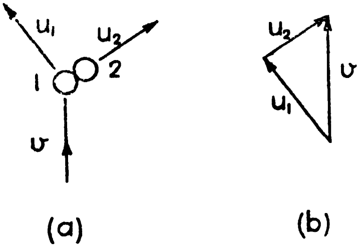


Fig. 2 (a). Billiard balls in oblique collision.
(b). Triangle of velocities in the same collision.

Before leaving the subject of billiard balls I should remark that our arguments cannot be applied without caution to what happens on a billiard table, since in that case the balls are rolling, i.e. they are rotating as well as moving. This complicates the problem a little and may cause slight changes in the answer.

ANGULAR MOMENTUM

Another important idea in mechanics which we shall want to use is angular momentum. For this, one must first explain what is meant by a moment. Moments appear in a simple way in the mechanics of rigid bodies, although otherwise rigid bodies will not be of interest to us.

It is a familiar fact that to turn a rigid body about a hinge the same force will be more effective if it acts further away from the hinge.

An important example of this is the lever, shown in Figure 3, which shows two weights *A* and *B* suspended by strings from points on a rod distant *a* and *b*, respectively, from the hinge, *O*. It is known then that the forces balance, i.e. have no tendency to turn the lever either way, if the weight of *A* times the distance *a* from the hinge of the point from which it is suspended equals the corresponding product *Bb* for the other. Therefore the force times the distance is a measure of the effectiveness of a force to turn a body about a fixed axis. For this product one uses the term "moment". This method of comparing different weights is used in one kind of weighing machine.

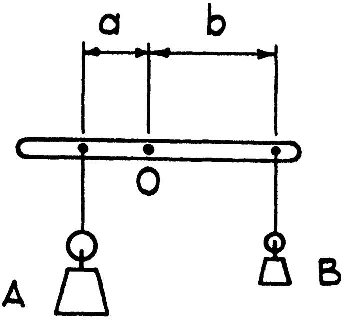


Fig. 3. Lever.

The case shown in Figure 3 is special because both weights pull on the horizontal lever in a vertical direction, i.e. at right angles to the lever. A force acting in some other direction can always be composed of a part at right angles to the lever and a part along the lever. It is evident that the part which pulls along the lever has no tendency to turn it about the axis. Figure 4 shows an arrangement in which the string supporting the weight runs over a pulley in line with the lever. It is clear that the force exerted by the string on the lever does not tend to turn it. We therefore define the moment of a force about an axis as the distance from the axis to the point of

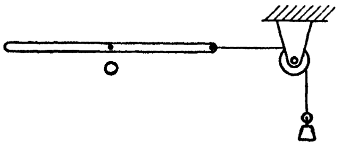


Fig. 4. No moment.

action, times the transverse part of the force. The transverse part means the part at right angles to the distance from the axis.

This is again illustrated in Figure 5, where the arrow *F* indicates the magnitude and direction of the force acting on some body at the point *P*, a distance *a* from the axis *O*. The force is shown by the triangle rule to consist of a transverse force F_t at right angles to *OP* and a longitudinal force F_l. Then the moment of this force about *O* is aF_t.

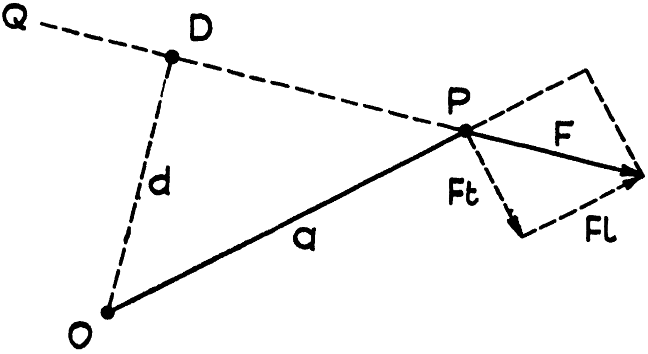


Fig. 5. Moments in general.

The same figure shows also another way of expressing the same moment. Consider the broken line *PQ* which is the line through *P* in the direction of the force. This is usually called the "line of action" of the force *F*. We have drawn the perpendicular from *O* on that line, which meets the line at *D;* this perpendicular is of length *d*. The moment of the force is then also equal to *Fd*.

That this is so will be evident to those readers who remember a little geometry, because the triangle *ODP* is similar to the triangle of forces and therefore *d* is in the same ratio to *a* as F_t is to *F*. Other readers will, I hope, accept the result that the two expressions for the moment amount to the same thing. Hence we may say that the moment of a force about an axis is the strength of the force times the distance of its line of action from the axis.

Having defined the moment of a force we define in a similar way the moment of momentum, or as it is more commonly called the "angular momentum" of an object. This is constructed like the moment of a force, but replacing the force by the momentum. We express our definition in the following way. The angular momentum of a small object *P* about an axis *O* is ap_t where *a* is the distance of the object from the axis, p its momentum (which we remember to be mass times velocity), p_t that part of the momentum which is in the direction at right angles to OP.

As for the moment of a force we can also use the alternative form which makes the angular momentum equal to dp where d is now the distance, from the axis O, of the line of motion of the object, i.e. the straight line through P in the direction of motion.

Using these definitions, we can then deduce from Newton's second law that the rate of change of the angular momentum of an object about an axis is equal to the moment of the force on the object about the same axis.

If we are dealing with more than one object and if some parts of our system exert forces on each other, then these forces do not affect the total angular momentum, just as we had found previously that they could be ignored in considering the conservation of momentum. This can be seen from Figure 6. P and Q are two objects which exert forces on each other. The Figure shows these forces to be attractive and therefore there is an arrow at P pointing towards Q and vice versa. They must be opposite and equal by Newton's third law. It is evident that these two forces have the same line of action and therefore the same d. They therefore have opposite and equal moments and therefore the angular momentum of P will, owing to this force, decrease at the same rate at which that of Q increases. The same is true if the forces are repulsive and if there are more than two objects.

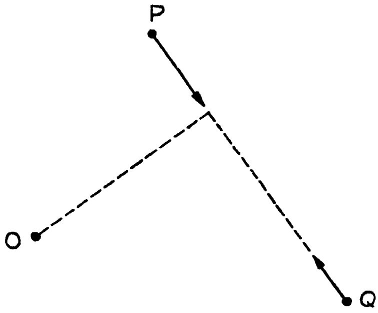


Fig. 6. Proving conservation of angular momentum.

This leads us to the law of conservation of angular momentum: if in a mechanical system which is free to rotate about an axis the only forces are those exerted by parts of the system on each other, then the angular momentum of the system about the axis remains constant. This is still true if there are external forces acting on the system, provided these external forces have no moment about the axis, i.e. they either act on the axis itself or if at any other point they act in a direction towards, or away from the axis.

Incidentally we have by this reasoning deduced from Newton's law the rule about equilibrium of a lever, which we had taken as the basis of our definitions, because for the lever not to start moving, the moments of all forces must balance. Since the force at the hinge has no moment about it, $aA-bB$ must be zero.

The law of conservation of angular momentum is a useful tool in solving many dynamical problems. It is of particular importance in problems involving bodies spinning about their axis, as for example with a spinning top, or in the mechanics of the game of bowls, but these applications are hard to explain without the use of mathematics and in any case are outside the scope of this book. Conservation of angular momentum is also of use in describing the motion of an object in a field of force, e.g. a planet around the sun. In fact it offers a simple way of showing the stability of such a motion. Suppose that at some instant the planet is at some distance r from the sun and moving at a velocity v in a transverse direction, i.e. in a direction at right angles to its distance from the sun. Then its angular momentum about the sun is mrv by the first definition. Now in this case angular momentum must be constant, since the attraction of the sun is directed towards it and therefore gives no moment about the sun.

Supposing that later the planet reached a point much closer to the sun, say a distance a from it. Then by conservation of angular momentum we have $mvr = mu_t a$ where u_t is the transverse part of its velocity at the second point. In other words $u_t \quad \frac{vr}{a}$ and since the total velocity cannot be less than its transverse part the velocity of the planet must be greater than its velocity at the first point, at least in the ratio $\frac{r}{a}$. As the planet gets closer and closer to the sun its velocity would therefore have to increase at least inversely as the distance, and its kinetic energy, which contains the square of the velocity, therefore at least as the inverse square of the distance. This, however, is incompatible with the conservation of energy. If we work out the potential energy due to the attraction of the sun, i.e. the work done on the planet by the force of gravity as it is moved from one point to another, we find that the potential energy, which in this case is negative, varies only as the inverse distance. In other words conservation of energy shows that the planet cannot, at a very small distance from the sun, have as much kinetic energy as the conservation of angular momentum would require it to have if it could get there, and it therefore follows that in the course of its motion the planet will not approach very closely to the sun and in particular will never fall into it. This holds as long as encounters with other planets, or friction with matter in inter-planetary space, have not produced forces whose moments will alter the angular momentum appreciably.

Another way of expressing the same fact is to say that as the planet gets closer to the sun its transverse velocity increases and with it the centrifugal force that would oppose its motion in a circle about the sun. Since the attractive force does not increase at a comparable rate the planet will in fact not move in a circle, but the curvature of its orbit will be less than that of a circle about the sun and therefore its distance from the sun will increase again.

STARTING CONDITIONS AND DEGREES OF FREEDOM

There is one other feature of the laws of mechanics to which I want to draw attention. Newton's laws allow us to express the acceleration, i.e. the rate of change of velocity of any object in terms of the forces acting on it. They do not tell us where to look for the object, nor what its speed or direction of motion is. Only if we know at some starting point the position of the object and its velocity at that point, then the laws determine how its velocity will change from there on, or how its velocity had been changing previously. Then the whole of its motion is determined. A complete statement of a problem in mechanics therefore includes two kinds of information: one is the general laws, i.e. Newton's laws together with a statement of what forces are acting and how they depend on the position and the velocity of the object. This information is usually called the "equations of motion". To this we must add, for example, some knowledge of its position and motion at the beginning and this is usually referred to as "initial conditions". Alternatively mechanical problems are sometimes specified in terms of some knowledge of how the motion starts and some information about where it is to finish. For example, we may ask in what direction a given gun has to be fired from a given point so that the bullet should hit a specified target. In that case we know the starting point and the magnitude of the initial velocity. We do not specify the initial direction, instead we specify a point which the bullet should reach. This together with the equations of motion again determines the problem completely. Such information which does not all refer to the starting point is more generally called "boundary conditions". The distinction between equations of motion and boundary conditions is one that recurs in all branches of physics.

We can then ask how many such data we must specify in order to determine our problem. In the case of a moving body three numbers are required to specify its position. This is the same thing as saying that space has three dimensions. For example we could specify a point near the earth by its geographical longitude and latitude, and the height above sea level. Its velocity also needs three parts to determine it. It could be stated for example by building it up from some velocity in the east-west direction, a north-south component and a vertical velocity. Or alternatively we could state its resultant velocity or speed and the direction in which it moves; a direction in space is specified by two numbers, for example the angle it forms with the vertical and the direction of the vertical plane in which the motion takes place.

If our mechanical problem is concerned with the motion of n objects we need $3n$ numbers to specify their positions and another $3n$ for their motion. But it may be that some of these objects are connected in a rigid manner so that they cannot move independently. This must always be the result of some forces acting between them, but in practice the forces which hold, for example, the different parts of a steel rod together, act always in such a way that the rod retains practically the same shape. As long as we are not interested in the internal forces in the rod we need not consider the mechanical laws for each part of the rod independently. In that case to specify the position of a rigid object completely we need, in fact, six numbers, as we can convince ourselves in this way: first specify the position of one end point of the rod, which requires three data. The position of the other end point would require another three, if it were not that we know the length of the rod and thus know beforehand that the other point must be at a fixed distance from the first. This leaves only two independent data to be specified (e.g. the direction in space of the axis of the rod which requires the knowledge of two angles). Having thus fixed the end points we are still free to turn the rod about the line joining them and to specify which way up the rod appears completely needs another number therefore. One therefore says that a rigid object, such as a steel rod, has six "degrees of freedom"; to specify its initial motion would require another six pieces of information and with those twelve initial data and the equations of motion the subsequent (or previous) history of the rod could be worked out.

The science of mechanics of which we have given a brief summary in this chapter was the first branch of natural science which could be reduced to a few simple basic laws from which everything else could then be derived if necessary by mathematical arguments. The success of its many applications to astronomy, to ballistics, to many engineering problems, left no doubt as to the soundness of the basic laws. This success of mechanics had the result that when new natural phenomena were studied one was inclined to describe them or explain them in terms of mechanical models. As we shall see this is possible in some instances (for example in the study of heat), but not in all, and we now realize that for a complete study of inanimate nature we must add to the basic laws of mechanics further laws of nature, some of which we now know fairly well, whereas others are still to be discovered.

2

Electricity and Magnetism

COULOMB'S LAW

THE description of electric and magnetic phenomena puts a greater strain on the imagination than that of mechanics. In mechanics we talk about the motion of objects, which is something we can see, and about forces, which we can feel. Electricity and magnetism usually have no direct effect on our senses. Admittedly we can get a pronounced sensation of electricity by putting two fingers on the terminals of the electricity mains, but this is hardly a convenient method of studying the characteristics of the electric circuit.

However, in this age there will hardly be a reader who is not familiar with at least some of the manifestations of electricity, or who would not be able to connect up an electric lamp or replace a battery. We may therefore take the existence, if not all the characteristics, of electricity and magnetism for granted.

One of the oldest known facts about electricity is that you can electrify rods of glass, rubber and other substances by rubbing them, and that they will then attract small pieces of very light substances, sometimes sufficiently to lift them up against their weight. A modern version is the parlour trick of rubbing a toy rubber balloon on a woollen garment and show it to stick to the ceiling. We now know that by rubbing we have caused the rods to accumulate an electric charge and that two such charges exert forces on each other which are similar to the gravitational force between two masses. Whereas, however, gravitational forces are always attractive, the forces between electric charges are sometimes repulsive, sometimes attractive. By comparing the forces between different charged bodies one can find that there are two kinds of charges, which are given the name "positive" and "negative", and that like charges repel and unlike charges attract each other. The magnitude of the force follows the inverse-square law as in the case of gravitation, and it is also proportional to the product of the quantities of charge in either body.

Quantitatively we have therefore a repulsive force of a magnitude given by Coulomb's law $\alpha \frac{qQ}{r^2}$ where Q and q are the two charges, r is their distance and α is a constant. Since we have not so far defined the quantity of electric charge, except through this law, the unit of charge can be chosen arbitrarily and we can therefore arrange for α to be 1. In other words, the unit of charge is that charge which on an equal charge at the distance of one centimetre will exert a force of one unit. (The unit here is the dyne, i.e. the force which will increase the velocity of a mass of 1 gm. by 1 cm. per second in a second.) Coulomb's law also incorporates the sign of the force correctly; if Q and q are both positive or both negative, their product is positive so that we get a positive force. If one is negative and the other positive, their product is negative and we obtain a negative repulsion, which is an attraction.

A similar law applies in magnetism for the interaction between two magnetic "poles". We find usually at the ends of a magnetized steel rod what is called a magnetic pole. These again are of two kinds, which one calls north and south poles, according to which of them will point towards the north if the magnet is freely suspended like a compass needle. As before two like poles repel each other, whereas opposite poles have a mutual attraction. The magnitude of this again is given by Coulomb's law with the pole strengths replacing the charges.

Magnetism differs, however, from electricity by the fact that we can never completely separate the poles. We cannot produce a north pole in a piece of iron without producing somewhere else in the same piece a south pole. If a magnetized knitting needle is broken in half then each half acquires a new pole at the break, so that each half again has a north pole and a south pole of the full strength. The reason for this will become clear later.

So far I have described electric forces by referring to the two charges which exert such a force on each other. In practical problems one is interested in the force on one such charge, without necessarily wanting to know whether this is due to one other charge nearby, or perhaps to a much stronger charge further away, or to a large number of different charges spread out somehow in the neighbourhood. It is therefore convenient to express the force on a small charged object as qE where q is the charge of the object (positive or negative) and E is called the electric intensity or electric field strength. E is a directed quantity, because it must indicate the direction as well as the strength of the force.

If, besides the object we are considering, there is only one other charge present we have from Coulomb's law $E = \frac{Q}{r^2}$ where Q is the magnitude of that other charge and r is the distance. The direction of this electric field is, of course, along the line joining the two charges. If instead we have many charges, then we have to regard E as the superposition of the field strengths due to every one of them.

E will, of course, depend on the position of the test object on which it acts. Now in any system of charges we can determine the field strength E at any point in the neighbourhood, regardless of whether at this point there is actually a charged object on which this field exerts a force. E is then that force which *would* act on an object carrying unit charge if it *were* placed in that particular position.

Similarly we speak of a magnetic field intensity, usually called H, and the magnetic force on a pole of strength p would then be pH.

As long as we are dealing only with electrostatics or magnetostatics, i.e. with the fields of charges and poles which are standing still, then the forces acting on a test object are conservative forces, in the sense in which we have defined this term in the last chapter. This means that if we move the test object from some point A to another point B, the force does a certain amount of work on the object. If no other forces act on it its kinetic energy has increased by that amount. If the object then moves back to the point A whether by the same path or in some other manner, the work done by the forces will be opposite and equal to the first, so that the total work done in moving from A through B back to A will be zero. Or, we may say, the force has no "circulation", in the sense in which we defined this term in Chapter 1.

That this statement is correct can be seen in the following way: If the force is due to a single charge we have an inverse-square law exactly as in Newton's law of gravitation, which we discussed in the previous chapter. This we had already seen to be a conservative field. If there is more than one charge present, the work done on the test object is composed of the work done by each charge. Since after a complete circuit none of the forces due to any of the charges has done any work, the combined field of all of them must be conservative.

This leads to the important conclusion that in the static electric field there always exists a potential, which we call V, and which varies from place to place. This potential is defined in such a way that if an object with charge q moves from a point A to a point B the work done by the electric force on the object is $q(V^A - V^B)$; in other words the bracket represents the difference between the potentials at A and at B.

Since the work was the force times distance it is easy to see that if we know the potential we can find the electric field strength as the rate of change of the potential with distance, i.e. the potential difference between two adjacent points divided by their distance.

In the case of the rods of glass or rubber, which we took as our starting-point, it is found that once a charge has been placed on some part of their surface it will remain there. There are, however, other materials known as conductors of electricity. These include all metals, and in a smaller degree many other solid substances, solutions of salts in water, etc. Such conductors can carry electric charge. If two opposite and equal charges are placed on opposite ends of such a conductor they will move towards each other because of their attraction and ultimately join up at the same place. The substance will then be uncharged because the effects of the opposite charges cancel each other. In fact we know that such conductors contain always both positive and negative charges, of which some at least are free to move. (We now know that it is the negative charges which are mobile in a metal.)

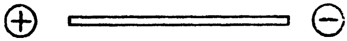


Fig. 7. Wire between two charges.

If such a conductor is placed in an electric field, as for instance in Figure 7. which shows a wire placed between a positive and a negative charge, the electric field of those charges then attracts the positive charges in the wire towards the right and the negative ones to the left. In other words one end of the wire now acquires a net positive, the other a net negative charge. This in turn affects the electric field in the neighbourhood. The movement of charges in the wire must continue until there remains no further electric field along the wire which could lead to a further movement of charges inside it. This final state has been reached when the electric potential is the same all over the wire. If there were a potential difference between any two points on the wire there would have to be also an electric field between those points, which would cause further movement of charges. It follows therefore that in a static field, i.e. when no charges are moving, the potential over any one conductor must be uniform.

For the same reason, if a certain quantity of charge is placed on a conductor its different parts repel each other and therefore move as far away from each other as possible, until the electric potential on the conductor is again everywhere the same. This means that the charge will end up on the surface of the conductor only, since any charge inside the conductor would have around it an electric field, which would therefore lead to a varying potential. The distribution of the charge over the surface of the conductor depends on its shape; in the case of a spherical conductor it is uniform all over. This explains the famous experiment of the "Faraday cage" which showed that the interior of a box of any shape with metallic walls is entirely free from electric forces, no matter what charges or electric fields act on the outside of the box.

Electric potentials can be measured directly with a device known as an electroscope. The simplest form of this consists of two very thin leaves, *L*, of metal foil suspended from a metal rod, which can be brought into contact with the conductor whose potential is to be measured (Figure 8). They are usually enclosed in a glass container to protect them from the effect of draughts. Suppose that the plate *P* of the electroscope is brought into contact with some object whose potential is higher (i.e. more positive) than that of the surroundings, then from the plate *P* and from the metal rod and the leaves an electric field must extend downwards and this would cause a variation of potential along the rod and along the leaves, unless some charge redistributed itself to make the potential constant. As a result the leaves become charged, and since they now both carry positive charge they must repel each other. They therefore rise from the vertical position to something like the angle shown in the figure. This angle is determined by the strength of their repulsion in relation to their weight, which tries to keep them vertical. By observing the angle one can therefore deduce the potential.

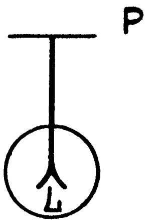


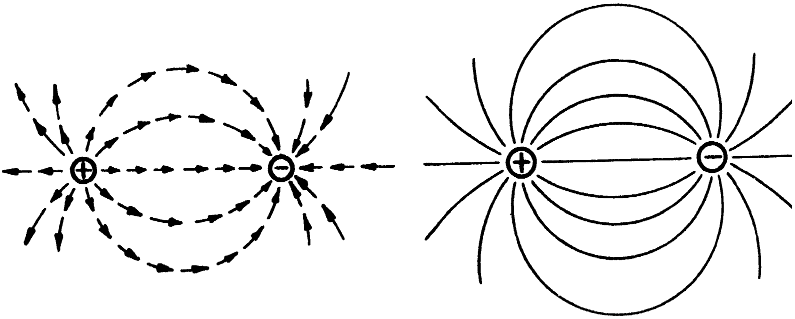
Fig. 8. Electroscope.

LINES OF FORCE

We have now stated completely the laws of electrostatics as applied to any system of conductors of arbitrary shape, with any charges on them; but it is not always easy to solve such a static problem, i.e. to determine just how to distribute a surface density of electric charge over the various conductors so as to make the potentials of each uniform. In solving such problems and also in illustrating graphically the nature of an electric field, one is greatly assisted by the concept of lines of force.

I have remarked earlier that the electric field intensity can be defined at any point in space and has direction as well as strength.

We may therefore illustrate the nature of an electric field by drawing at many points in space little arrows, the direction of each arrow indicating the direction of the electric field strength at that point. An example of this has been sketched in Figure 9a for the case of two equal and opposite charges. Near the positive charge the arrows point away from it because here a positively charged test object would be repelled from the charge. Near the negative charge they point towards it, and at other places the resultant electric field can be obtained by working out the resultant of the two fields due to the individual charges, each of these fields being given by Coulomb's law. Now a line of force is a line drawn by joining up a succession of these arrows in such a way that the line has at any point the direction of the arrow, i.e. of the electric field at that point. This has been carried out in Figure 9b. Readers have probably seen this situation illustrated by spreading iron filings near the north and south pole of a permanent magnet. The iron filings always take the direction of the magnetic field; while the result is of the same kind as Figure 9a the eye tends to join up the pattern into the continuous lines of Figure 9b.



(a) (b)

Fig. 9 (a). Field marked out by arrows.
(b). Lines of force.

It would appear that in this way lines of force may serve to illustrate the direction, but not the strength of the field. However, it turns out that if we draw such a pattern it also determines the variation of the field strength. The field is strongest at the points where lines of force crowd together and weakest where they are widely spaced. In other words the density of lines of force, i.e. the number that intersect a small area of a given size placed at right angles to their direction is a measure of the field intensity.

This statement is evidently correct for the field of a single point charge. Consider Figure 10. The charge is at Q; the lines of force are straight lines emanating from Q, since the force is everywhere in the direction away from Q. Consider the bundle of lines of force which pass through the small ring A_1 whose distance from Q is r_1; if we follow them out to distance r_2 we need a larger ring of area A_2 to encircle them. It is evident from the figure that the diameters of the rings are in the proportion r_2 to r_1. Or in other words if r_2 is twice as large as r_1, then the diameter of the ring A_2 is twice as large as that of A_1. The area of a ring varies as the square of its diameter and therefore the area A_2 is then four times that of A_1. This means that the area occupied by the same bundle of lines of force is at a distance r_2 four times as large as at r_1 and the number of lines of force per unit area is four times smaller. More generally the number of lines of force per unit area therefore varies as the inverse square of the distance from Q and therefore varies in the same manner as the electric field intensity varies by Coulomb's law.

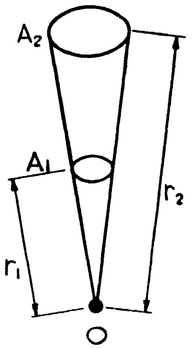


Fig. 10. Lines of force and Coulomb's law.

That this statement is true not merely for a single charge but for any general combination of electric charges cannot be demonstrated quite as easily, and I hope the reader will accept it as a statement of a plausible fact. It is usually referred to as the theorem of Gauss.

Lines of force drawn in this manner will always start from a positive charge and end on a negative charge, or in the case of an open system not enclosed in any metallic box, some come in from infinite distance (as in the case of a single negative charge) or go out to infinite distance (as in the case of a single positive charge).

The total number of lines of force starting from any given positive charge of amount Q is easily seen to be $4\pi Q$. This follows because if we draw a sphere of radius r around the charge its surface area is $4\pi r^2$. The number of lines of force per unit area must equal the electric field intensity at this distance from the charge, which is $\frac{Q}{r^2}$ and the total number must therefore come to $4\pi Q$.

We have therefore seen that a static electric field has two important characteristics. Firstly, it can be described by lines of force, both as regards direction and magnitude. A field of this kind is called "divergence-free". Strictly speaking a divergence-free field is one in which lines of force do not start or end anywhere. We must therefore supplement the statement by adding that on any positive (negative) charge of amount Q, $4\pi Q$ lines of force must begin (end). Secondly the electric field gives rise to a conservative field of force. Such fields are cometimes called "irrotational" since no work is done in taking a test charge around any closed circuit returning to the starting-point, and therefore the force has no tendency of making such test objects go round in circles. An irrotational field is a field without circulation.

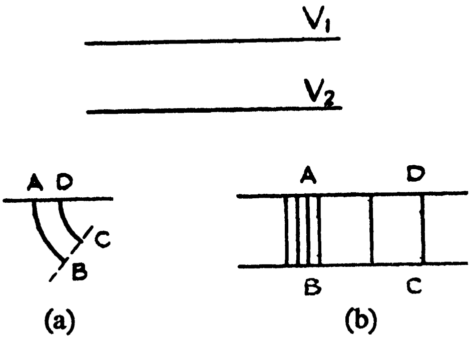


Fig. 11. Parallel-plate condenser:
(a). Lines of force must be straight.
(b). Lines of force must be equally spaced.

Now the important point is that these two statements of the field being divergence-free and irrotational can completely replace Coulomb's law. If we know where the lines of force start and end, and if we know that they give rise to an irrotational field, their whole course is uniquely determined.

How this is used to work out charge distributions and fields in practice is illustrated by the simple example of Figure 11. This shows a section through two parallel metal plates, a distance d apart, which are kept at different potentials V_1 and V_2. Such an arrangement is known as a "condenser". Consider the picture of lines of force in this case. There are no charges present except on the upper plate, which we assume to be the positive one, and negative charges on the lower. Therefore lines of force will run from the upper plate to the lower. If the width of the plates is large compared with the distance, as in the figure, in fact nearly all the lines will start from the bottom of plate 1 and go to the top of plate 2. They must start perpendicularly from the one plate and arrive perpendicularly at the other. This follows because the potential on each plate must be constant, and therefore there must be no part of the electric field strength in the direction of the plate. The field strength must therefore be directed at right angles to it.

Now assume that after starting from the plate the lines of force tended to bend, for example as in Figure 11a, and consider two adjacent lines of force starting from the plate at A and D and both turning. Then if we moved a test charge from A to B along one line of force the electric field would do work on it. Moving across to the other line of force, from B to C, no work will be done because were are moving at right angles to the direction of the line of force and therefore at right angles to the electric force. The component of the field in the direction of motion is therefore zero. Next, we move our test body from C to D along the second line of force and here we move in a direction opposite to the force, so our test charge does work against the field. But the distance between the lines of force which we have shown remains approximately constant, i.e. the field strength is constant, but we have moved a shorter distance on the return journey, and therefore the work lost is less than the work gained on the outward journey. Finally from D back to A we are at constant potentials so no work is done. Hence the curved lines of force of Figure 11a would necessarily have circulation.

In this manner, other possible shapes of the lines can be ruled out and we conclude that everywhere except near the edge of the plates the lines of force must be straight. Their density must also be the same everywhere, for if they were denser at one point than at another as in Figure 11b, then there would a circulation along the closed path $ABCDA$. Hence the lines of force are straight and equally spaced. The field strength is everywhere vertical and of the same magnitude. In that case the work done in going from plate 1 to plate 2 is Ed where E is the field strength and this must equal $V_2 - V_1$, therefore the field strength is $E = \frac{V_2 - V_1}{d}$.

The number of lines per unit area starting in the one plate is E and the amount of charge per unit area is therefore $\frac{E}{4\pi}$ or $\frac{V_2 - V_1}{4\pi d}$.

This is the correct result for the charge per unit area on the plate of a condenser. It could have been obtained without the concept of lines of force and without the rules about the irrotational and divergence-free nature of the field, but this would have involved rather more mathematics.

FIELD LAWS OR ACTION AT A DISTANCE

I have gone through this example in so much detail in order to illustrate how these characteristics of the field can be used for mapping it out. But there is more to it than convenience in solving problems. We have replaced Coulomb's law by the law that the field is divergence-free and irrotational and that lines of force originate and terminate on charges. By doing this we have replaced a law about action at a distance by a law concerning the change of the field from point to point. In formulating basic laws of physics there has always been a tendency to avoid the idea that the presence of a body in one place could have a direct influence on another body somewhere else. Take as a simple illustration a mechanical door-bell. By pulling a handle at the door we can make a bell ring in another part of the house. We are not satisfied with describing this as a force exerted by the handle on the bell, until we have understood that there is a wire connecting the two, and that the handle merely exerts a force on the wire attached to it. Each piece of wire, on being pulled, acquires a tension and this pulls on the adjacent piece of the wire and so it goes on until the pull has been transmitted to the bell at the other end.

The relevant laws of physics are the laws of mechanics and the laws of elasticity, the latter saying that a piece of wire will oppose by a strong force any attempt to stretch it. Both the mechanical laws and the laws of elasticity therefore are local laws concerning the behaviour of the wire at any point.

The difference between this situation and a real action at a distance becomes quite important when we consider a wire of great length, as in the case of the wire operating a signal along a railway track. Then the slight extension of the wire under the effect of the tension applied to it is cumulative and as a result the signal lever at the other end does not start moving immediately when the signalman begins to apply force; owing to its elasticity the wire transmits the action only with a finite speed.

The discovery of electricity presented a different situation, in that apparently, as Coulomb's law showed, a charge at one point could be influenced directly by the presence of another charge somewhere else. It is therefore an obvious idea to ask whether there might not also be some agency which would transmit this action. At first one tried to invoke the idea of a mechanical medium which might transmit such actions, and it was called the "ether".

The ether had to have some very peculiar qualities for this purpose. It had to fill all space, since electric forces can be transmitted from any point of space to another. It had equally to pass right through all material bodies, since electric actions can also be transmitted through all solid bodies which are not conductors of electricity. If it was an elastic medium which could transmit electric disturbances as tensions and displacements, then it would have to be extremely stiff to permit the transmission of actions with great speed, and at the same time it had to permit the passage of solid bodies through it without any noticeable resistance or friction.

Such an ether would be most unlike any elastic substance within our experience. The attempt to explain electric forces in such a way was therefore not attractive. It was finally reduced to an absurdity when it was recognized that the laws of elasticity are not basic laws of nature. As we shall see in later chapters the behaviour of elastic bodies can be understood in terms of the atoms which compose them and the forces between those atoms. We shall also see that these inter-atomic forces are, in the last resort, due to electric forces, and if we therefore try to explain electric forces in terms of elasticity we would merely be involved in a vicious circle.

But while physicists were thus led to abandon the idea of a mechanical medium they could still maintain the idea of local laws. The laws concern the behaviour of the electric field and link the field at one point in space to that at adjacent points. What this electric field is we cannot explain any more than we can explain the concepts of mass or force. Each represents something abstracted from our experience, and a concept that we have found necessary in describing the laws of nature. The difference is that mass and force come from experience directly accessible to our senses, whereas the concept of an electric field comes from the experience of the scientist in his search for the laws of nature. It is not surprising that the one should at first sight appear more natural and more familiar than the other.

Up to now we have been concerned only with electrostatics or magnetostatics, i.e. with charges or magnets at rest. At this stage the distinction between field equations and the idea of action at a distance is purely a matter of choice. We have found laws for the electric field, which are the complete equivalent of Coulomb's law, and we may use one or the other. Later, however, we shall see that, as in the example of operating a signal by pulling a wire, so also in the electric case the action is transmitted with a finite, though very large, speed, and in order to give account of this we must use the field concept.

ELECTROMAGNETISM

The first extension beyond the stage of static fields is to consider electric currents. We have already recognized the possibility of an electric charge moving along a conductor, but up to now we have only considered the final result when the flow of charge had ceased. While electricity is moving along a conductor it represents an electric current. Its strength is measured by the amount of electric charge which is being carried along the conductor per second. We can obtain such currents for short periods by connecting, by means of a metal wire, two conductors which are originally at different potentials, for example the two plates of a condenser as in Figure 11. But since the supply of charge is limited in such a case the current will not last very long. Steadier currents can be obtained from an electrolytic cell or "battery", whose principle we shall not discuss. It is sufficient for us to note that as a result of chemical action in the battery it has a tendency to set up a certain difference in potential between its two terminals.

In this way and in many other ways it is possible to maintain a steady flow of electricity through a conductor for a reasonable period. One then finds that to maintain a steady current a steady difference of potential between the ends of the wire is required. There is then inside the wire an electric field which exerts a force on charges inside the wire. In such a field the charges would therefore be continuously accelerated and their velocity would increase, unless there was some friction or resistance opposing their movement. All normal metals have such a finite resistance, which is always opposed to the motion of the charges and therefore represents a non-conservative force. In the flow of electricity through conductors of finite resistance energy is converted into heat and this fact is exploited in electric heaters or electric lamps of the filament type. In most other applications of electricity it is a necessary evil that one tries to minimize, like the effect of friction in machines.

As soon as we experiment with currents in conductors we notice a new phenomenon, namely that an electric current also gives rise to a magnetic field. Observation (e.g. by placing a compass needle close to a wire carrying a current) shows that the magnetic field of a current in a long straight wire is as shown in Figure 12. The wire is taken to be at right angles to the plane of the paper, so that the black circle in the middle represents a section of the wire. The lines of force are circles around the wire. The intensity of the field falls off inversely as the distance r from it. The magnetic field is proportional to the strength of the current in the wire, and we can write it in the form $H = \frac{2i}{rc}$ where i is the strength of the current in the wire (i.e. the amount of charge passing through it per unit time), and c is a new constant, which can be determined experimentally. c turns out to be 30,000,000,000 centimetres per second. When one handles numbers as large as this it is not convenient to have to write down that many zeros and a convenient abbreviation use in such cases is to write $c = 3 \times 10^{10}$ cm/sec. This is reasonable since the square of 10 is 100, i.e. a figure 1 with two zeros, the third power of 10 is a 1 and three zeros and similarly the tenth power of ten is a 1 followed by 10 zeros.

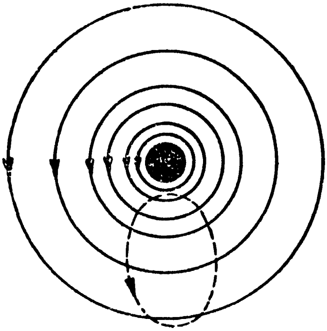


Fig. 12. Magnetic field around a current.

This magnetic field is evidently no longer irrotational. If we took a magnetic pole around one of the concentric circles we would be following a line of force, and the magnetic force on the pole would always be in the direction of motion, therefore the force would be doing work on the pole all the way. Since the work is the force times the distance one can easily see that the circulation on a circle of radius r around the wire is $\frac{4\pi i}{c}$ since the circumference of the circle is $2\pi r$. The result is therefore the same, no matter which of the circles is followed, and indeed it can be shown to be the same even if we do not go along an exact circle, but follow any other path that encloses the wire. The circulation of the magnetic force around a wire carrying a current is $\frac{4\pi i}{c}$.

One can also show that there is no circulation if the loop does not encircle the wire, such as the one shown by the broken line in Figure 12. The reason is that in this case the pole has of necessity to travel some part of the way with the field and some part of the way against the field. If the direction of the field and the direction of travel of the pole are as indicated by the arrows it will travel for a longer way with the field than against it, but where it travels against the field, on the inside of the loop, the field is more intense. Therefore for the whole space outside of the wire the field still satisfies the simple laws we had found previously. It is divergence free and it is irrotational. Only on the wire itself the field is not irrotational, so that the second of our laws must there be changed.

What we have learnt so far about the laws of the electric and magnetic field can be summarized in this way: at any place where there are no charges and no electric current the electric and the magnetic fields are irrotational and divergence free, i.e. they can be described by lines of force which do not start or stop and have no circulation. In a region where there are *charges* the *electric* field is no longer divergence free; the number of lines of force which leave the region minus the number which enter equals 4π times the net charge in it, i.e. the difference between positive and negative charge. In the presence of an electric *current* the *magnetic* field is no longer irrotational. The magnetic circulation around a loop equals $\frac{4\pi}{c}$ times the current passing through the loop.

In this form the laws are valid provided the fields do not change with time. This is the case if we do not have electric charges accumulating anywhere in space, or spreading out, and if any currents are steady.

We note in particular that the magnetic field remains always divergence free, that is to say that magnetic lines of force never start or end anywhere. They either form closed loops or come in from infinite distance and go away into infinite distance. This would not be true if there existed really magnetic poles, because on these magnetic lines of force would have to start or to end. However, we have never seen any free magnetic poles and all characteristics of magnets (made for example of steel) can be explained, as we shall see later, in terms of small loops of current inside the steel. We have to picture the lines of force which come from the north pole of a bar magnet and return to the south pole, closed up inside the magnet by going back from the south pole to the north pole. Figure 13a shows the external lines of force from a bar magnet. Figure 13b shows a coil of wire carrying a current which has the same external lines of force, but with the lines closed up by returning through the inside of the coil. We must not, however, picture a magnet as a single coiled-up current, but rather as a large number of very small loops of current distributed throughout the body of the steel.

The magnetic effect of a current is of the greatest practical importance, because it means that an electric current exerts a force on a magnet, or on a substance such as iron, which becomes magnetized in the presence of a magnetic field. This is used in a great many practical devices and it also offers the most convenient method of measuring the strength of an electric current.

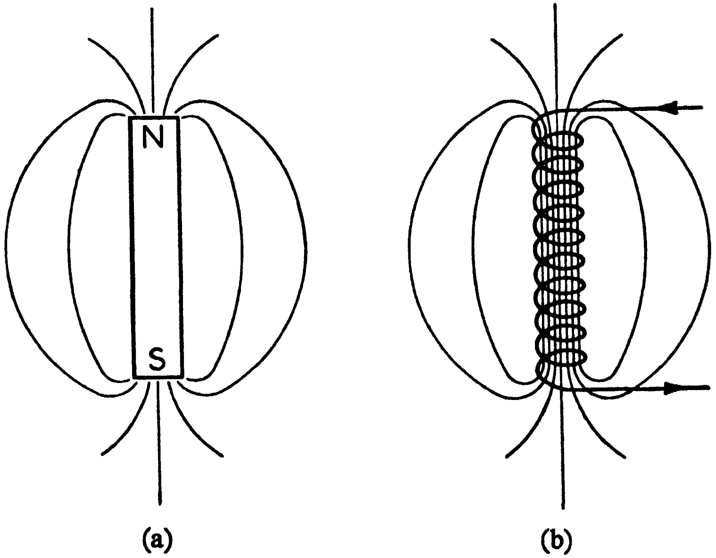


Fig. 13 (a). Bar magnet.
(b). Magnetic field of a coil.

A device for measuring an electric current in this way is known as a galvanometer. In its simplest form the current is passed through a coil of wire in order to increase its magnetic effect, and a magnetic needle is suspended like a compass needle either inside or near the coil. The deflection of the needle then is a measure of the strength of the magnetic field, hence of the current.

It is in fact so much easier to measure a current in this manner than by finding directly the charge it has carried in a certain time, that one often regards this magnetic effect as the definition of the strength of a current. This led to the so-called "electromagnetic" unit for measuring a current. Since in the formula for the magnetic field the current appears in the combination $\frac{i}{c}$ one has therefore sometimes chosen to take $\frac{i}{c}$ as a quantity indicating the strength of the current. We shall in this book make no use of such electromagnetic units. I have mentioned them only because readers who are familiar with electricity may remember the laws in the form based on electromagnetic units and they might be puzzled by the slightly different appearance of our results.

From the law of action and reaction it follows that if a current exerts a force on a magnet, a magnet must also exert a force on a current. In other words, there must be a force on a current exposed to a magnetic field. This argument is quite sufficient to determine the strength and the direction of this force. We shall not go through the reasoning in detail, but merely state the result: the force on a wire of length l carrying a current of strength i in a magnetic field of strength H is at right angles to the field and at right angles to the direction of the current. Its strength is $F = \frac{ilH_t}{c}$ where H_t is the transverse component of the magnetic field, i.e. the part of the magnetic field strength in the direction at right angles to the wire.

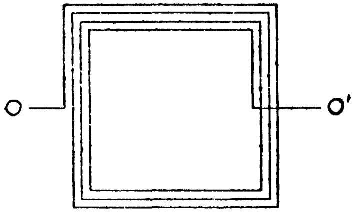


Fig. 14. Principle of electric motor.

This force on a conductor in a magnetic field is used in the electric motor. Consider, for example, the arrangement sketched in Figure 14. Many turns of a wire carrying an electric current are wound on a rectangular frame, which is free to rotate about the axis OO′. A uniform magnetic field is provided, which we assume to be in the vertical direction. Then if the current is in the direction from O to O′, the force on the wires in the upper edge of the frame will be at right angles to the plane of the drawing and for the sake of argument we shall take it to be forward. At the bottom of the frame the current flows in the opposite direction and the force will therefore push backwards. The net result is therefore a tendency to make the frame rotate about the axis OO′. This continues until the frame has turned through a right angle, when the force no longer assists in turning the frame. If by its inertia the frame goes on rotating the edge carrying the current from left to right would be below the other edge and the force would now try to turn the frame in the opposite direction. For this reason the motor is provided with a device known as a "commutator". This is a system of sliding contacts by means of which the direction of the current is reversed as the frame passes through the horizontal position. When the frame is again vertical the current therefore again flows from left to right on the upper edge of the frame, and the force continues to pull in the same direction. The magnetic field which is necessary for the functioning of the motor may be provided either by permanent magnet or by a fixed coil carrying an electric current.

In practice a motor does not consist of a single rotating coil, but of very many of them and contains an iron core which enhances the magnetic effects.

INDUCTION. GENERATOR AND TRANSFORMER

From these practical illustrations we return to the basic laws. New phenomena appear when we consider fields or currents which are not steady, but which change with time, either because the source which supplies the electric current alters with time or because a magnet or a wire carrying a current is moving. The most important new fact then arising was discovered by Faraday. If a closed loop of wire is placed in a magnetic field, and if either the magnetic field changes with time or the loop of wire is moved, then a current will flow round the loop. We had seen that to cause a current to flow along a wire, i.e. to move the electric charges contained in the wire, there must be an electric field present along the wire. Observation shows that in the case of a circular loop of wire the electric field is of strength $\frac{1}{2\pi rc}$ times the rate of change of the "magnetic flux", i.e. the rate of change of the number of magnetic lines of force passing through the loop. This answer is correct when the loop is standing still and the magnetic field changes, perhaps because a magnet nearby is moving or because we are switching on or off the current in the coil nearby. It is also correct when the loop of wire is moving in a steady field.

The second case is of particular importance for the electric generator, which is built just like the electric motor of Figure 14. If the rectangular coil of Figure 14 is made to turn about the axis by a force from outside, the number of magnetic lines of force that pass through it will keep changing. It will be greatest when the frame is horizontal, and zero when it is vertical. When it is horizontal again, but facing the opposite way, the magnetic lines of force will pass the loop of wire in the opposite direction. Therefore the total number of lines of force, or "flux", through it varies from a greatest positive amount for one horizontal position to an equal and opposite amount for the other horizontal position and back again. Its rate of change is greatest in the vertical positions and the sliding contact of the commutator will connect the coil with the external circuit just when it is nearly in the vertical position. The rotating coil will therefore supply a pulse which tries to drive current through the external circuit always in the same direction. By using very many coils connected at different angles there will always be one at any given instant which acts in the right direction.

However, for the purpose of getting a simple statement of the fundamental laws we are more interested in the situation when the loop of wire is standing in a changing magnetic field. We then see that the electric field which is driving the current round the loop is different in kind from any electric field we have met before, because previously the electric field always was irrotational, i.e. no work would be done on a test charge which describes a closed loop in the field. Now we have seen that the field acts always in the same sense around the loop and the work done on a test charge is then equal to the distance it had to go round the loop, i.e. the circumference $2\pi r$ of the circle times the strength of the electric field at each point, and this comes to $\frac{1}{c}$ times the rate of change with time of the magnetic flux.

This conclusion turns out to be independent of the shape of the loop. We can therefore account for the phenomenon of electromagnetic induction by amending our previous law about the electric field to read: the electric field is irrotational in the absence of changing magnetic fields; otherwise its circulation round a loop equals $\frac{1}{c}$ times the rate of change of magnetic flux through the loop.

The electric effect of a changing magnetic field is also exploited in many important practical applications. One of them is the transformer, of which the principle is illustrated in Figure 15. Two coils of wire A and B are placed so closely together that the magnetic lines of force produced by A (or at least some of them) pass through the coil B. A current passes through A from the end a_1 to a_2 and this current is made to reverse direction rapidly, going through a complete cycle n times per second. (The "frequency" n for current taken from the electricity mains is usually 50 or 60 cycles per second.) As this current changes its direction the magnetic field which it produces changes and hence also the flux through coil B. This, as we have seen, causes an electric field along that coil, which also changes direction n times per second and which will cause a current to flow if the terminals b_1 and b_2 are connected through a circuit, which may in general contain also a resistance.

The current through the external circuit is the same as if one applied to the terminals b_1 and b_2 an oscillating potential difference whose magnitude is again given by the electric field along the coil, times the total length of wire in the coil. One therefore sometimes talks of the potential difference or "voltage" generated by the transformer. However, one must be careful with the use of this term, since we have seen that the idea of potential is applicable only to an irrotational electric field. In fact the usual elementary introduction to electricity, which starts at the beginning from the concept of potential, may leave the student in a confused state of mind when he asks himself how the potential varies along a circuit which contains a transformer coil or an electric generator. To avoid this difficulty one uses the term "electromotive force" (e.m.f.) for the electric field times the length of the wire, or what amounts to the same, to the equivalent potential difference one would have to apply to the terminals b_1, b_2 to obtain the same current in the external circuit.

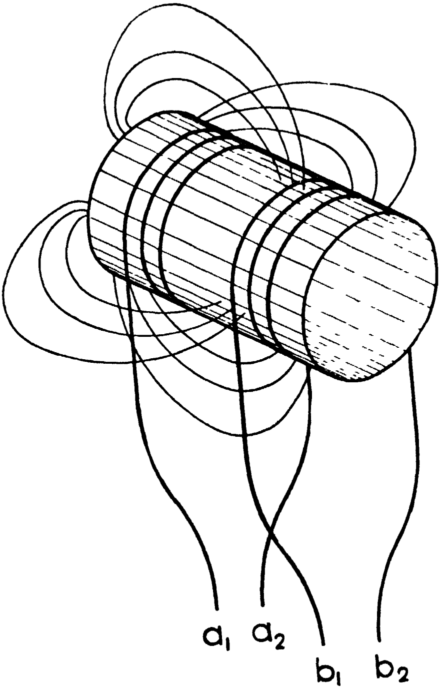


Fig. 15. Principle of transformer.

If the current and the form of the primary coil *A* is given and therefore its magnetic field and the rate at which it varies, then the electromotive force in the "secondary" coil *B* is proportional to the number of turns in it. By making the number of turns in the secondary coil very large we can therefore obtain a very high electromotive force, but on the other hand we can then only take a very weak current from the secondary circuit, because if this current were too strong it would, flowing through many turns, produce a strong magnetic field which would react back on the primary coil *A*. Such transformers are therefore used to obtain from a given alternating current in another circuit a current which is very much weaker, but has a very much higher equivalent voltage and therefore can flow more easily through high resistances. Or conversely by making the number of turns in the secondary coil *B* less than that in the primary, the transformer can be used to set a current flowing whose intensity is much greater than the maximum current we could take out of the primary circuit.

The same phenomenon of induction can also be of importance if we deal only with a single coil. In Figure 16 we have a circuit formed by connecting the coil *C* through a switch *S* with the terminals of the battery *B*. If the switch is closed a current will flow through the coil and therefore produces a magnetic field whose lines of force are as sketched in Figure 13b. They therefore pass through the loops of the coil. If the circuit is broken by opening the switch the current must stop and therefore the magnetic field will disappear. Hence the flux through the coil *C* changes by a large amount in a very short time, and the law of induction therefore requires that a very strong electric field is set up along the wire of the coil. This produces a very strong electromotive force, usually sufficient to cause charge to flow through the air gap of the switch *S*, resulting in a visible spark. The induction of electric forces in a coil due to the magnetic field of the current in the coil itself is known as "self-induction" and a circuit very much like that of Figure 16 is used to provide the spark for the ignition in the most common type of motor-car engine.

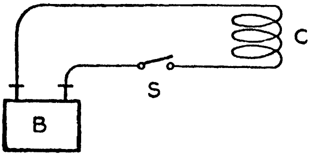
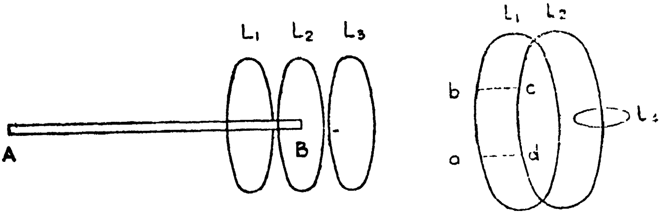


Fig. 16. Self-inductance.

There is another instructive way of looking at this situation. As we shall see later, both electric and magnetic fields contain stored energy, and a coil like that of Figure 16, which produces an intense magnetic field, therefore contains a large amount of magnetic energy when the current is flowing. Because of the law of conservation of energy this stored energy cannot be destroyed, but only transformed into other forms of energy. As we break the circuit work must be done on some other part of the system. The visible spark is a very clear indication of where the work is being done; the light from the spark indicates that some part of the air has got very hot and the process may be summarized by saying that we have converted the magnetic energy from the coil into heat. The magnetic energy of such a coil therefore causes a tendency for the current to continue just as the kinetic energy of a large fly-wheel means that the wheel has a strong tendency to keep rotating and that something drastic is bound to happen if we force it to stop.

COMPLETION OF THE LAWS

We return now to our description of the fundamental laws. As we have them so far they are adequate to deal with any situation normally met by the electrical engineer, provided he is not concerned with extremely high frequencies. They can describe adequately what happens in ordinary electric circuits, say at mains frequency, whether there are generators, motors, transformers, etc. However, it was discovered by Clerk Maxwell, who is responsible for formulating these laws clearly in a complete form, that as they stand now they are not quite consistent, and that one can derive a contradiction from them. To show this consider the situation shown in Figure 17. We imagine that an alternating current flows along the straight wire *AB*: when the current flows from *A* to *B* it carries charge from left to right. We have therefore positive charge accumulating at *B* and negative charge at *A*. This cannot go on indefinitely because this charge will set up an electric field trying to reverse the current, but we can imagine that the current flows for a short time in this direction and then reverses and goes on alternating in this manner. We do not have to consider at the moment exactly how this alternating current is produced.



(a) (b)

Fig. 17 (a). Alternating current in a wire.
(b). An apparent contradiction.

Now consider the magnetic field produced by this current. From our present laws we expect to find a magnetic field whose lines of force are circles about the axis of the wire. Indeed imagine a loop drawn in space like L_1 then there must be a magnetic force acting round this loop, since an electric current is passing through the plane of the loop. We are not so sure what happens in the case of the loop L_2, because a plane through this loop just passes the end B of the wire, i.e. the point where the current stops. But if we go on to consider the loop L_3 there is no current through it and therefore we should expect no magnetic force around it. We would therefore conclude from our present wording of the laws that there should be a magnetic force acting around loop L_1, but not around loop L_3. This, however, is not permissible. To see this, consider Figure 17b, which shows the loops L_1 and L_3 enlarged. We can imagine a closed loop in space *abcda* and we may ask for the magnetic circulation around this loop. As we move a pole from *a* to *b* a magnetic force acts on it. No work is done between *b* and *c*, or between *d* and *a*, because here it moves in a direction parallel to the wire, and the magnetic lines of force of the field due to the current are always in planes perpendicular to the wire. Lastly no work is done between *c* and *d* because we are here following the loop L_3, along which we expect no magnetic force. We conclude therefore that there is a circulation round the small loop *abcda*. But this contradicts our law for the magnetic field, since no current passes through this loop. This argument shows that there is still something missing in our fundamental laws. Maxwell showed that the only way to avoid this contradiction is to assume that there is a magnetic circulation, not only in the presence of an electric current, but also when there is a changing electric field.

This removes the discrepancy because in the arrangement of Figure 17 we have at any instant a charge at A and an opposite and equal charge at B. Say, for the sake of argument, that the charge at B is positive. Then lines of force will go from B to A in the manner shown in Figure 9b and some of these will go outwards through the loop L_3 and some will go back through loop L_1. Now at the time when the current is greatest the rate at which charge accumulates at the end of the wire is greatest and therefore the rate of change of the charges and hence of the electric field always follows the current. If a changing electric flux gives rise to a circulation in the magnetic field, then the magnetic circulation around the loop L_3 is no longer zero, the one around the loop L_1 is reduced, and there is also a magnetic force around the loop *abcda*.

To make things just right one has to postulate that the magnetic circulation around any loop is equal to the sum of two parts, the first part being, as before, equal to $\frac{4\pi}{c}$ times the total current through the loop, the second part being $\frac{1}{c}$ times the rate of change of the electric flux, i.e. of the number of electric lines of force passing through the loop.

We shall not give the quantitative reasoning which shows that this term just corrects the discrepancy, nor show that if we include this term in our laws they are now always consistent.

In discussing the statement that a changing electric flux produces the same magnetic effect as a current, Maxwell used the term "displacement current", which implies that a kind of current, not associated with the transport of any charge, flows in empty space wherever the field changes in time. This idea belongs to a picture of the electromagnetic field which has since become obsolete and we shall not use it.

We are now in a position to state the complete laws of the electromagnetic field as follows:

The magnetic field is always divergence free. The electric field is divergence free, except where there are charges, and the number of lines of force starting in a given volume minus the number terminating is given by the net charge in that volume. The electric field is irrotational except where there are changing magnetic fields. The circulation round a loop equals $\frac{1}{c}$ times the rate of change of the magnetic flux through the loop. The magnetic field is irrotational except where there are currents or changing electric fields. The circulation round a loop is $\frac{4\pi}{c}$ times the current passing through the loop minus $\frac{1}{c}$ times the rate of change of the electric flux through it.

The new part which we have added to the laws is unimportant as long as we are dealing with low frequencies and systems of a moderate size, because for the intensities of the electric field with which we are normally concerned and their usual rate of change, the change of electric flux is in practice so small that the intensity of the magnetic field caused by it is negligible.

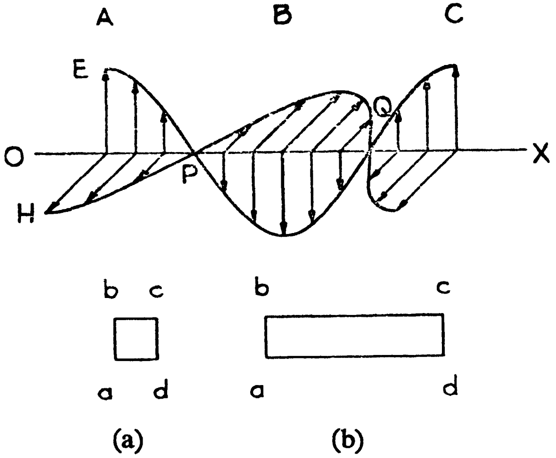
However, this is different at high frequencies, and what happens then can be understood by returning to Figure 17a. We have already seen that the magnetic field must now extend beyond the ends of the wire because there must be a magnetic force around the loop L_3. This of course is an alternating magnetic force, since in our example the current and everything with it changes direction many times a second. Hence along the loop L_3 there is a changing magnetic field and this in turn gives rise to an electric field by the law of induction. For example if we consider the small circular loop l_4 in Figure 17b it now encircles a line along which the magnetic flux is changing and around it we have an electric field acting. In following the electromagnetic lines of force we find that this electric force along the loop l_4 cannot be entirely accounted for by the electric lines of force which connect A with B. There must be in addition further lines of force, and since there are no charges except at A and B these must be closed loops.

For this reason the electric field extends beyond the magnetic loop L_3, and according to our new law the changing electric field must produce magnetic fields further away from the ends of the wire. These in turn produce electric fields by the law of induction, and as a result we find that electric and magnetic fields extend further out into space than they would in the static case. We cannot by our clumsy methods, which avoid mathematical description, decide easily the form which the fields will take, but a complete mathematical treatment leads to the result that they are waves. We can at least verify that such waves are in agreement with our fundamental laws, i.e. with Maxwell's equations.

ELECTROMAGNETIC WAVES

A typical such wave is illustrated in Figure 18. The wave travels in the direction of OX, and at any instant the electric field is the same, i.e. has the same direction and the same magnitude on each plane at right angles to this line. In other words if we choose any point on the line OX the electric field at that point is the same as at any point above or in front or behind it, and the same is true of the magnetic field. We need therefore indicate only how the electric and magnetic fields vary along this line. This has been done by showing arrows to indicate the field at some instant. As shown the electrical field is always vertical, but it is upward at some points and downwards in others. The magnetic field is at right angles both to the electric field and to the direction of travel; it is sometimes forward and sometimes backwards. As time goes on the whole pattern travels in the direction from O to X.

We now fix our attention on some point P, and we consider a small rectangular loop of height l near this point as shown in Figure 18a. There is now an electric force acting around this loop in a clockwise direction, because the electric field acting upwards along the left-hand edge is stronger than that on the right-hand edge and there are no horizontal electric fields on the other edges. Hence if we take a charge up from *a* to *b* work is done which is not completely cancelled by the negative work done as it returns from *c* to *d*. This result is in agreement with the basic laws, because there is also an increasing magnetic flux through this loop, since the direction of the magnetic field is from the rear towards the front of the drawing and its intensity increases with time as the whole pattern shifts towards the right. At a point such as *Q* the magnetic flux is decreasing since this point has nearly been reached by the "node" at which the magnetic field vanishes. But in this case also the electric field increases from left to right so that the electric force round a similar loop is in the anti-clockwise direction. By drawing small horizontal rectangles we can equally verify that the magnetic force round a closed loop is in agreement with the rate of change of the electric flux through the loop, as should be the case. We thus see that such a wave is qualitatively of the right form to agree with our laws. The agreement will be quantitatively correct, however, only if the velocity with which the wave travels is correct.



(a) (b)

Fig. 18. Electromagnetic wave:
(a). Near point *P*.
(b). From *A* to *B*.

To see this we introduce the "wavelength" λ, which is the distance after which the wave form repeats itself completely. This means the distance from *A* to *C*. Consider now a vertical rectangular loop whose length is half the wavelength, i.e. which extends from A to B; its height is again l. Then the electric circulation around such a loop is $2El$, where E is the maximum electric field strength, since the work on a unit electric charge is El on going up the left-hand side and again El on going down the right-hand side of the loop. This should equal $\frac{1}{c}$ times the rate of change of the magnetic flux through the loop. This magnetic flux is zero at the instant we have drawn, since the magnetic flux is forward through one half and backwards through the other half of the loop. However, it keeps changing. For example if we take an instant of time earlier than the one shown, by $\frac{1}{4}t$, where t is the time taken to travel one wavelength, the points A and B were two nodes of the magnetic field and therefore the average magnetic field between A and B was H_m, the mean magnetic field between two nodes. At that instant the flux through our loop was therefore H_m times the area of the loop, i.e. $\frac{1}{2}\lambda l H_m$. On the other hand, an instant $\frac{1}{4}t$ after the one shown, A and B are again nodes, but the magnetic field is in the opposite direction. We may therefore say that in a time interval of length $\frac{1}{2}t$ the flux through the loop has changed by $\lambda l H_m$, and therefore its average rate of change over this time interval is $\frac{\lambda l H_m}{\frac{1}{2}t}$.

Now we are not concerned with the *average* rate of change over such a time interval, but with the *maximum* rate of change, because the instant for which we have drawn the wave is the instant at which the rate of change of flux through the given loop is greatest. The variation with time of this rate of change follows the same pattern as the magnetic field and therefore we obtain the *maximum* rate of change if we take in our equation the *maximum* field H instead of the mean field H_m. Equating therefore the electric force round the loop with $\frac{1}{c}$ times the rate of change of the magnetic flux we find the result $2E = \frac{\lambda H}{\frac{1}{2}ct}$ or $E = \frac{\lambda H}{ct}$.

Now we had defined t as the time during which the wave travels one wavelength λ and therefore $\frac{\lambda}{t}$ is the velocity v of the wave. We have therefore $E = \frac{v}{c} H$.

In just the same way we can apply the other law, which relates the magnetic circulation round a loop to the change of electric flux, and the answer comes out to be $H = \frac{v}{c} E$. In other words E and H are seen from the one law to be in the same proportion as v to c, from the other in the proportion of c to v. Clearly this is possible only if $v = c$, and H and E are equal.

We have therefore derived the result that in an electromagnetic wave the magnetic and electric fields strengths are equal, that they are directed at right angles to each other and at right angles to the direction of propagation and that the wave travels with the velocity c.

Soon after Maxwell had derived these conclusions from the laws of the electromagnetic field, H. Hertz proved the existence of such waves by experiment and therefore established confidence in Maxwell's deductions. Such electromagnetic waves are used as radio waves to transmit signals, speech and pictures. Our primitive Figure 17 contains the principle of the simplest radio aerial. In practice one has to add to this some device which will actually cause electric current of high frequency to flow along the wire.

Our argument also shows that electromagnetic waves are transverse waves, i.e. that the electric and magnetic fields must be at right angles to the direction of travel. Indeed if we tried to vary our figure by having one or both of the fields parallel to the direction OX we would not be able to obtain agreement with the basic laws.

If we consider such a transverse wave we may ask for its state of "polarization". For example in the waves shown in Figure 18 the electric field is always in a vertical direction. We could equally well have considered a wave travelling along OX but with its electric field horizontal and its magnetic field in the vertical direction. These can be distinguished. For example a radio receiver with its aerial straight and vertical would respond to one but not the other. Another possible form of wave is obtained by combining two waves of the same wavelength and the same direction of propagation, one with its electric field vertical, the other with it horizontal, and following each other in such a way that the nodes of the one follow a quarter wavelength behind those of the other. Consider then a point such as A in Figure 18, at which at this instant the one wave has its electric field upwards. When this wave has travelled a quarter wavelength its electric field at this point will have dropped to zero, but by that time that of the other wave will have reached its maximum and will, let us say, point forward. After another quarter wave the node of the horizontal wave will have arrived and there will be no horizontal part to the electric field at this point, but by now the first wave contributes an electric field in the downward direction and so it goes on. One can show that in such a case the electric field is always of constant magnitude, but its direction revolves at constant speed about the axis of propagation turning from the upwards, through the forward and to the downward direction and so on.

This is what we call a wave with "right-handed circular polarization". In a left-handed wave the electric field revolves in the opposite direction. Just as we can obtain a wave of circular polarization by combining two waves of plane polarization as in Figure 18, we conversely construct a plane polarized wave by suitably combining two waves of opposite circular polarization.

FIELD ENERGY

One last remark about the electromagnetic laws concerns the question of energy. Consider two electric charges at a finite distance from each other. They repel each other, and therefore in bringing them together from large distances we have done work against this repulsive force. We can recover this work by allowing the charges to separate again. As in all mechanical problems we think of such recoverable energy as stored in the form of potential energy. As long as we are concerned only with such a static problem it does not matter whether we imagine that this potential energy is stored in the charges themselves or distributed over the field, but we have much less choice when it comes to wave phenomena. This follows because waves may transmit energy from a radio transmitter to a receiver. As their speed is limited the energy will have been lost by the transmitter before it is picked up by the receiver, and if energy is to be conserved it must have been somewhere else in the meantime. This can only be in the wave which is travelling from one to the other. We must therefore think of the energy as residing in the field. We can determine the amount of field energy by working out how much energy has been lost by the charges and currents which have produced the field and allow for any work the field may have done on other charges or currents.

We shall not follow this argument through in quantitative detail, but state the answer, which is that the energy density, i.e. the energy per unit volume, is equal to $\frac{1}{8\pi}(E^2 + H^2)$.

That this is a reasonable form for the energy can be seen by applying it to the particular example of a parallel-plate condenser, Figure 11. We had seen that in this case the electric field between the plates is uniform and of magnitude $E = \frac{V_2 - V_1}{d}$. The volume of the space between the plates is dA where A is the area of one plate, and the total energy, which is the energy per unit volume times the volume, therefore is $\frac{(V_1-V_2)^2A}{8\pi d}$. Now we had also seen that the charge per unit area is $\frac{V_1-V_2}{4\pi d}$ and therefore the total charge is $Q=\frac{(V_1-V_2)A}{4\pi d}$. Our expression for the energy therefore implies that the total energy is $\frac{1}{2}(V_1-V_2)Q$. At first sight one might think that the energy should be twice this amount, because we can imagine the final state of the condenser reached by starting with the plates neutral and then move charge from one plate to the other. Since the energy required for this is the quantity of charge times the potential difference, the energy would appear to be $Q(V_1-V_2)$. However, we must remember that when the plates are uncharged they also have no potential difference and therefore, if we move the charges over gradually, we start by moving charge against zero potential difference, and the potential difference rises from nothing to the amount V_1-V_2. The average potential difference against which we work is therefore just $\frac{1}{2}(V_1-V_2)$ and this shows that our expression for the energy is correct.

That any magnetic field must equally be associated with an energy density became clear already in our discussion of self-inductance.

Readers who have some experience with electricity and magnetism may be surprised that throughout this chapter on electromagnetism we have not mentioned the concept of dielectric constant and magnetic permeability, or the electric induction, D, and the magnetic induction B. The explanation is that these do not appear in the fundamental laws; they are required only to describe the behaviour of material objects in electric and magnetic fields. This behaviour can now be in principle explained in terms of their atomic structure, and from this point of view these concepts are merely convenient mathematical constructions which are helpful in describing how systems containing large numbers of atoms react to an electric or magnetic field. They are consequences of the basic laws of physics and not part of them.

3

Light

SPECTRUM. INTERFERENCE. WAVES

NEWTON, who laid the foundations of the science of mechanics, was also very interested in the properties of light. He studied amongst other things the phenomenon of the spectrum, i.e. the coloured band which is obtained when a narrow beam of white light passes through a piece of glass or some other transparent object, whose faces are not parallel. The best-known example of this is the rainbow, caused when light from the sun passes through raindrops in the air. The commonest way of looking at this phenomenon in the laboratory is to use a prism, i.e. a piece of glass with a triangular section, as in Figure 20.

In either case we see instead of white light a band showing "all the colours of the rainbow", starting with red at the one end and changing through orange, yellow, green to blue and finally purple.

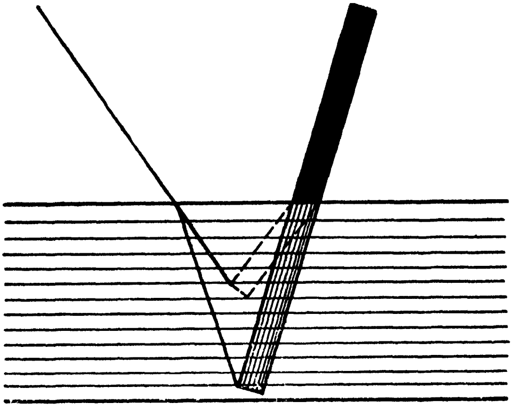


Fig. 19. Refraction.

Newton recognized that this means that white light is actually a mixture of all these colours and that the water drop or prism is merely a means of separating the different colours from each other.

This is due to the phenomenon of refraction, i.e. to the fact that when light passes obliquely through a surface separating two different transparent substances, such as glass and air, or water and air, as in our examples, the light ray is broken and changes direction. This can most easily be demonstrated by the observation that a straight stick which is placed with one end in water and which is not perpendicular to the water surface, appears to have a kink at the point where it enters the water (see Figure 19). The "spectrum" occurs because the light rays of different colours in the white light are refracted by different amounts. This has been indicated in Figure 20 where the full lines are meant to indicate the path of red light and the broken lines that of purple light.

We shall not here go into the explanation for refraction, since this has to do with the atomic structure of the transparent substance. But it is important to realize that we can expect to find simple laws for light only if we look, not at ordinary white light, but at light of a pure colour. We may obtain this, for example, by passing light through a prism, Figure 20, and then put in its way a plate with a narrow slit which lets through only a narrow part of the spectrum, of a particular colour. Such light of a pure colour is called monochromatic light and a device to pick a pure colour out of the ordinary white light is known as a monochromator.

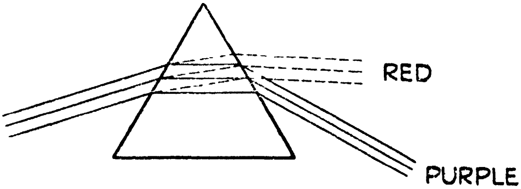


Fig. 20. Prism.

A simpler way of obtaining light of a pure colour is to use a source which gives directly such light. For this purpose we may take a gas flame which does not give much light, such as the bluish gas flame of an ordinary gas cooker, and place in it a substance containing sodium, for example ordinary cooking salt. This gives a bright yellow light, which turns out to be practically monochromatic. The same kind of light is also obtained from electric discharge tubes containing sodium, which are sometimes used as street lamps. The reason for this will become clear later.

Newton also observed the "interference" of light. A typical case of this consists of looking at a source of monochromatic light, e.g.

a sodium flame, reflected in a thin glass plate with very accurately plane faces. In this case we see the reflection of the flame not uniform, but with alternating bright and dark stripes, the so-called interference fringes, running across it. Newton did not find the explanation for interference, but this was brought out later by the work of Huygens and Fresnel. They showed that one would expect just such interference fringes to occur if light consisted of waves.

Waves are known in many different branches of physics, and they always occur when an extended medium is disturbed by some external force. The commonest example in our everyday experience is surface waves on water, which arise when the surface of the sea is disturbed by the wind or when we throw a stone into a pond. In the latter case they are usually of the form of a regular ripple, and a section through this at a place some distance away from where the stone was dropped might look like Figure 21. Here the straight horizontal line indicates the water surface before the ripple passed and the wavy line indicates the disturbance. It consists of wave crests following each other at regular intervals with depressions or "troughs" in between them. The distance between two successive wave crests is known as the wavelength, usually denoted by the Greek letter λ. The number of wave crests passing a fixed point in a second is called the frequency. It equals the wave velocity, divided by the wavelength. Another form of wave motion occurs in sound. When sound travels through air we know that this is also due to a wave in the air. If the sound is that of a pure musical note we have a simple wave like Figure 21, but in this case the curve represents, instead of the elevation of the water surface, the changes of pressure in the air. At the wave crests the air is compressed and in the troughs

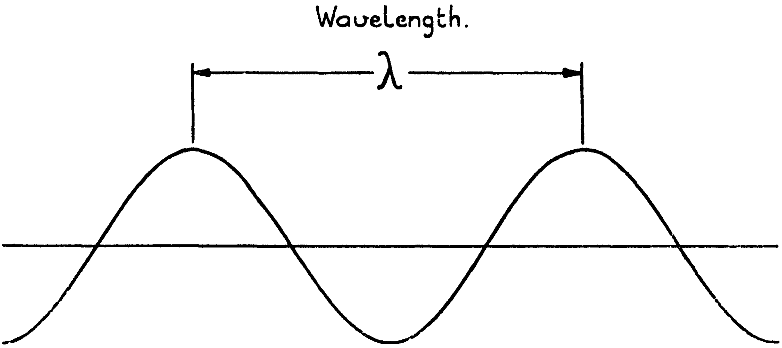


Fig. 21. Wave.

it is expanded, i.e. at less than normal pressure. As the wave passes, the wave crests and troughs move and therefore a place which at one time was a trough will have become a crest when half a wavelength has passed. This means the air at this point, which at first was expanded, has now become compressed and some air must have moved to this point. When another half wavelength has passed the same amount of air will have moved back again. It follows therefore that the movement of the air in this case is in the direction in which the wave travels, in the figure from left to right and back to left again. In such a case we speak of longitudinal waves. In the water wave on the other hand the water surface is rising up and dropping down again, and the motion is therefore mainly in a direction at right angles to the way in which the wave is travelling; we speak here of a transverse wave.

We have already met the case of electromagnetic waves in the previous chapter, and we shall see presently that these are very closely connected with light.

In order to discuss how interference comes about we need not ask for the nature of the wave, all that matters is that we have some quantity which oscillates in the manner sketched in Figure 21.

Consider such a wave reflected from a glass plate, as in Figure 22. Here the inclined lines marked S_1 and S_2 indicate the surfaces of the glass plate. The wave arrives from the left, and the shaded regions indicate the wave crests. Some of this light is reflected on the front surface of the glass plate. In the position illustrated in the figure, a crest is just passing over this surface and therefore a wave crest is beginning to be reflected upwards. When the incident wave was just one wavelength further back, the previous wave crest had passed the surface S_1 and given rise to another reflected wave crest, which since then has already travelled one wavelength upwards. This is now shown in the position c_1, the previous one as c_2, etc.

However, only a small part of the light is reflected at the surface of the glass plate and the rest travels on (actually in a slightly changed direction, because of the refraction, but we shall disregard this for simplicity), and it continues until it reaches the rear surface of the glass, marked S_2. In the conditions chosen for the figure a trough is just passing the rear surface at this instant and by the same reasoning as before the reflected wave passing upwards also has a trough in this place. The last wave crest passed at a time when the incident wave was half a wave further back and since then it has already travelled half a wavelength upwards.

In this way, we recognize that in the two reflected waves, one coming from the front, the other from the rear of the glass, the crests are staggered, those of the one wave being at the same height as the troughs of the other. These waves actually overlap. We have shown the incident wave and the two reflected waves as very narrow, since otherwise the drawing would have got confused, but in actual fact the incident wave has a considerable extension sideways and the same is true of the reflected waves, therefore the reflected waves travelling upwards actually coincide; and since at every place their effects are in opposite directions, the one producing a crest where the other produces a trough, the net result will be that they cancel each other out. In the situation shown in this figure we would therefore not observe any reflected light at all.

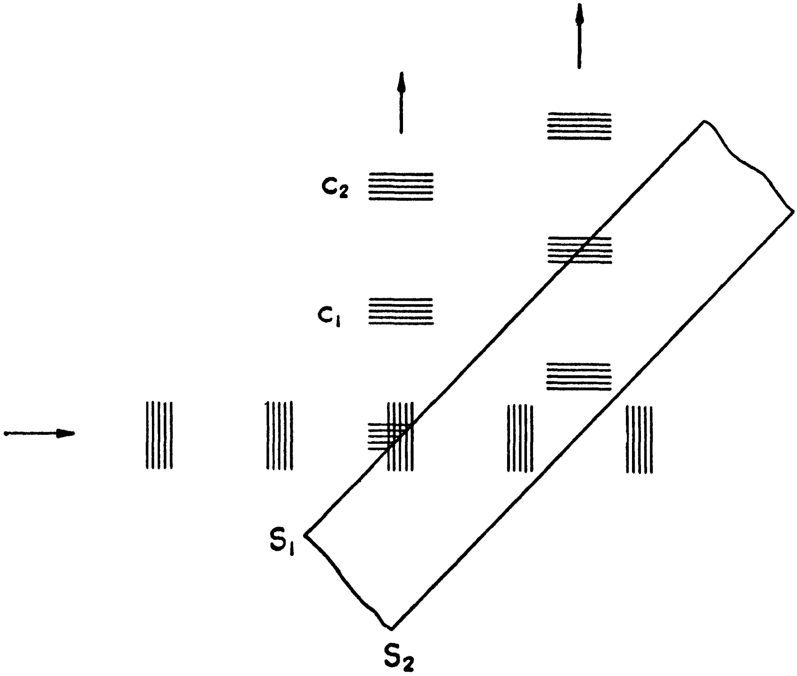


Fig. 22. Interference at a glass plate.

This result evidently depends on a precise comparison between the path of the wave across the plate and the wavelength. If the plate had been a little thicker we would have had a position when the arrival of a wave crest S_1 would coincide with the arrival of a wave crest at S_2. In this case the two reflected waves would have reinforced each other instead of cancelling. The result also depends on the angle at which the wave meets the plate. If we turn the plate more towards the vertical direction we shorten the path of the wave across the plate and there will be a position again when the waves reinforce each other. This is precisely the situation that we have in looking at a sodium flame reflected in a glass plate. We see the different parts of the flame reflected at slightly different angles in the plate, and if light consists of waves then we do indeed expect that at some angle we find a bright reflection and at other angles none.

All the details of such interference fringes, i.e. the way they depend on the thickness of the plate and on the angles and the way the intensity varies from a bright to a dark band, agree completely with the predictions of wave theory. By comparing the spacing of the fringes for a given angle and a given thickness of plate, we can find what the wavelength of the waves must be.

In this manner one finds that the wavelength of visible light is somewhat less than one ten-thousandth part of a centimetre. For example the wavelength of sodium light is about 6×10^{-5} cm. (Here we have used an extension of the abbreviation mentioned previously. By 10^{-5} we mean $\frac{1}{100{,}000}$, i.e. 1 divided by a 1 with five zeros, or in decimal form 0·00001, i.e. a decimal fraction in which 5 zeros occur.) We can now also understand what is the difference between the different colours. Each colour belongs to a particular wavelength. Red light has the longest waves, and throughout the rainbow the wavelengths decrease, the shortest wavelength, that of violet light, being roughly one-half that of red light.

If we look at interference, using not monochromatic light, but white light, then the waves from the red part of the spectrum will have their dark fringes in different places from those of, say, the blue light, and in that way we obtain coloured fringes which are not so easy to see. They are much clearer when instead of a glass plate we have a thin film, whose thickness is not very different from the light wavelength itself. In that case the fringes are much broader and the colours appear more clearly. This is the explanation for the colours of soap bubbles, or of oil films on water.

DIFFRACTION, POLARIZATION, SPEED

Another important way of demonstrating interference is the use of a diffraction grating. By this we mean a set of parallel lines spaced at equal distances. For example these could be lines scratched into the surface of a mirror, or fine wires stretched out parallel to each other. If light waves fall on such a diffraction grating we get a situation like that of Figure 23. Here the circles show the lines of the grating in section, and an incident wave is shown, whose direction of travel

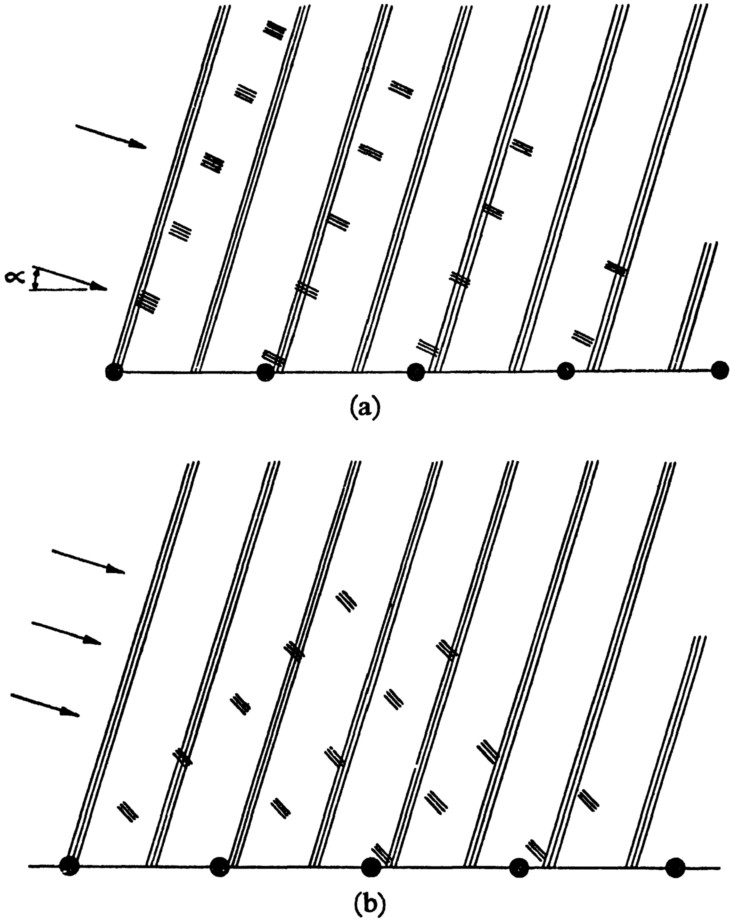


Fig. 23 (a). Diffraction from a grating.
(b). The same in a different direction.

is at an angle a to the plane of the grating. Again wave crests are shown shaded. We get a reflected wave from each line of the grating. The short shaded pieces show these reflected in some particular direction. All these reflected waves combine and the result will depend on whether the reflected waves from the different lines all have the crests in the same position and therefore reinforce each other, or whether they cancel. Figure 23a shows a situation where they cancel and 23b shows the same grating, for light of the same wavelength, but we consider reflected waves going off in a slightly different direction and now all the different waves reinforce each other.

The problem differs from that of interference fringes in a glass plate inasmuch as we combine not merely two but many reflected waves, and in general we have waves with all kinds of positions of their wave crests, which must therefore cancel out, except for very exceptional angles where all these waves have their wave crests in the same place, or as it is called, are in phase with each other. Such a diffraction grating therefore reflects monochromatic light only in very sharply defined directions, or will draw out white light into well defined coloured spectra.

If the spacing of the lines on the grating is much larger than the wavelength the sharpest lines are obtained if both the incident wave and the direction in which we look for the reflected light are almost parallel to the surface of the grating. This can be used to demonstrate the diffraction very easily without the use of any apparatus. We have only to look at the reflection of a fairly small light source, for example an ordinary electric light bulb, in a gramophone record, holding the record so that we see it almost edge-on. In this case the ridges of the record act as a diffraction grating and we see a number of images of the light, each drawn out into a coloured spectrum. This simple demonstration shows the principle of the diffraction spectrograph, an instrument by which physicists can determine the wavelength of light to a very high accuracy.

The fact that we have to look at the gramophone record almost edge-on shows that the wavelength of light is much smaller than the distance between the grooves of the gramophone record. These are usually about 1/40 of a centimetre apart and by estimating the angle at which the different diffraction spectra appear, it is quite easy to see that the wavelength of the light is a few hundred times smaller than the spacing, in agreement with the figure quoted before.

The study of light also reveals that it consists of transverse waves. This follows because we can produce two light waves of the same colour travelling in the same direction, which are nevertheless different in that the direction of the vibration in the wave is different. We speak then of differently "polarized" light. To-day the best-known form of this is obtained by means of "polaroid" filters. These are plastic sheets which are fairly well transparent for light polarized in a particular way, but opaque for light polarized at right angles to this direction. If two such filters are placed on top of each other and one turned relative to the other we can find a position where the pair lets through almost as much light as each filter separately. In this case the filters are in the same position. Turning one of the filters through a right angle stops practically all light coming through. This effect is exploited in one form of three-dimensional film, in which the illusion of depth is produced by making the two eyes of the spectator see slightly different images, as they do in looking at real objects in space. The spectator wears polaroid glasses so that the left eye sees only light whose polarization is vertical and the right eye that with horizontal polarization. The light projected on the cinema screen is made up of two images, each of polarized light and each photographed by a camera placed in the position of the left or the right eye respectively.

All these results show that light consists of transverse waves. Their velocity, which is very high, has been measured in several different ways.

The first determination of the finite speed of light came from an observation by the Danish astronomer Rømer, who had observed the path of the satellite (or moon) of the planet Jupiter when the earth was fairly close to Jupiter. He predicted from this that at a certain time later the satellite should pass in front of the planet itself and therefore its shadow should be visible. This did take place, but rather later than predicted. It seemed impossible to account for this discrepancy through any errors in the calculation or in the observation of the orbit of the satellite. Rømer saw that this must mean that the light took a finite time to travel from Jupiter to the earth; since Jupiter now was further away than at the beginning, the passage of the satellite would be seen here later for that reason. Rømer determined the speed of light that would have to be assumed to account for the discrepancy, and his result was not far from the correct figure, about 3×10^{10} cms. per second, which has since been confirmed by more accurate observations of the same type and also by laboratory experiment.

Now we had seen that electromagnetic waves have a speed of just this magnitude and they are also transverse waves. It is therefore natural to regard light waves as a form of electromagnetic waves.

LIGHT CONSISTS OF ELECTROMAGNETIC WAVES

This idea was fully borne out by further investigation. Some support to this idea comes at once from the fact that if light waves are in fact electromagnetic waves we should in general expect good conductors of electricity to be opaque for light. We expect this because an electromagnetic wave contains electric fields and these would set the mobile electric charges in a conductor into motion. Now we had seen that the motion of such charges is subject to friction or "resistance", and therefore as they move they will convert energy into heat. By the conservation of energy, this must be supplied from somewhere and it can only come from the energy carried by the electromagnetic wave. This must result in a weakening of this wave. Such a wave can therefore not travel far through a conductor before nearly all its energy has been converted into heat. In fact all metals do absorb light. A more careful study of what happens when an electromagnetic wave meets the surface of a conductor leads to the conclusion that most of the wave will in fact be reflected; it is again true that clean surfaces of metals act as reflectors of light, i.e. as mirrors.

The final confirmation that light consists of electromagnetic waves comes from the fact that our knowledge of the structure of matter then leads us to a complete understanding of the emission and absorption of light and also of the refraction and reflection of light by different materials. To this point we shall return in discussing atomic structure.

This is therefore one of the cases in which a whole branch of physics can be reduced to the basic laws of nature deduced from another branch. The laws of the electromagnetic field which were summarized in the preceding chapter contain all the fundamental laws required for the propagation of light. This does not mean that optics, or the science of light, is less important to-day than it was earlier, but it is now concerned with the application of the known basic laws, with the study of the response of various materials to light and with the design of optical instruments and the functioning of the human eye. It no longer contains any mysteries as regards the laws of nature, except for the quantum phenomena, to which we shall return in Chapter 7.

Light waves are then of the same kind as the electromagnetic waves which are used in radio communications. The only difference between these two cases is the difference in wavelength, light waves being about a million times shorter in wavelength than the shortest wave used, for example, for television broadcasts.

Electromagnetic waves of any wavelength between these two regions can also be produced, but they are not visible to the eye. Wavelengths from about one-hundredth of a centimetre down are contained in radiant heat. The heat we feel in placing our hand near to a hot stove is due to the fact that any hot object sends out heat waves, which are electromagnetic waves of a wavelength much longer than that of light, but shorter than that of radio waves.

Heat waves of rather shorter wavelength and only slightly longer than that of red light are known as "infra-red" radiation. These are produced with great intensity from hot objects which are nearly red hot, i.e. just not hot enough to produce visible light.

If an object, e.g. a metal wire, is heated further it appears red, and at a still higher temperature it becomes white hot, i.e. it produces light covering the whole visible spectrum. It then also begins to produce light of a wavelength too short to be visible, and such waves, whose position in the spectrum is just beyond the violet end, are known as ultra-violet light. Such ultra-violet light is particularly easily produced in an electric discharge tube, and a common form of such a discharge tube is used in the mercury vapour lamp, which gives quite intense ultra-violet radiation.

If we increase the potential difference between the ends of such a discharge tube further we can produce ultra-violet light of shorter and shorter wavelength. At wavelengths a few thousandths times that of visible light the waves have again different characteristics and are called X-rays; waves with a wavelength a few thousand times shorter still than that of X-rays are known as gamma-rays, and with these we shall again be concerned later.

The different names indicate merely that waves of such different ranges of wavelength differ in the sources by which they can be conveniently produced, in their effects on different substances, and in the methods by which they are conveniently detected. Nevertheless the same basic laws, i.e. the laws of the electromagnetic field, apply to them uniformly over the whole range from radio waves to gamma-rays.

It might therefore be said that in this chapter we have learnt nothing new about the laws of nature. We have merely found a new field of application for laws that we knew already. Nevertheless light and other electromagnetic radiation has proved, as we shall see, of the greatest importance in extending the laws of nature. Its study was vital both in the understanding of the structure of matter and in showing that refinements are required in the fundamental laws of mechanics and of the electromagnetic field.

EMISSION THEORY. GEOMETRICAL OPTICS

It is interesting to compare the wave theory of light with earlier ideas. The most obvious picture was that a lamp, or other source of light, was sending out small particles which travel with high velocity in a straight line until they reach our eye or a photographic plate, or some other device by which they can be detected. This would explain why light travels in straight lines. If there were suitable forces acting on these "light particles" whenever they pass the boundary between air and glass, or any two transparent media, one could explain the refraction. For example if we played a game of bowls on a bowling green that had two different levels, one-half of it being raised and the two levels joined by a short and fairly steep bank, each bowl starting from the upper level would describe a path which would look just like a light ray passing from air into glass. Its direction on the lower level would not be the same as that in which it arrived on the upper level, but it would make a larger angle with the bank. The "emission theory" of light is therefore capable of explaining the behaviour of light rays in passing through prisms and lenses. Also the fact that a prism separates colours could be understood if we believed that the force acting on the light particles at the surface had a different effect on those of blue light and those of red light.

The limitations of this emission theory are shown up by interference and diffraction, which clearly establish the wave nature of light.

However, the interference and diffraction effects are important only if we are concerned with distances which are not very large compared with the wavelength of light. We do not therefore in most practical cases notice these effects unless we specially go out of our way to do so. It is true that we have talked about making diffraction apparent by means of a gramophone record, although the distance of the grooves from each other is still very large compared to the wavelength. However, we had there to look at the record in a special way, namely at such a small angle that the grooves appeared to be only a few light wavelengths apart.

The wave theory of light gives us a much deeper insight into its nature, but it would not be satisfactory unless it also preserved the simple features of the emission theory wherever they are applicable, i.e. when we are concerned only with objects at distances very large compared to the light wavelength. It is important that the wave theory predicts for such a case precisely the same behaviour of light as would have followed from the simple emission theory.

In order to see how this comes about consider first of all the simplest situation, namely light which still consists of plane waves, in other words is extended very far at right angles to the direction in which it travels, but only lasts a short time. This holds for example, for the flash from a flash bulb a long distance away.

If we draw a line in the direction in which the light travels and at some instant consider the electric field intensity along this line we obtain a picture like Figure 24a. *OX* is the line along which we

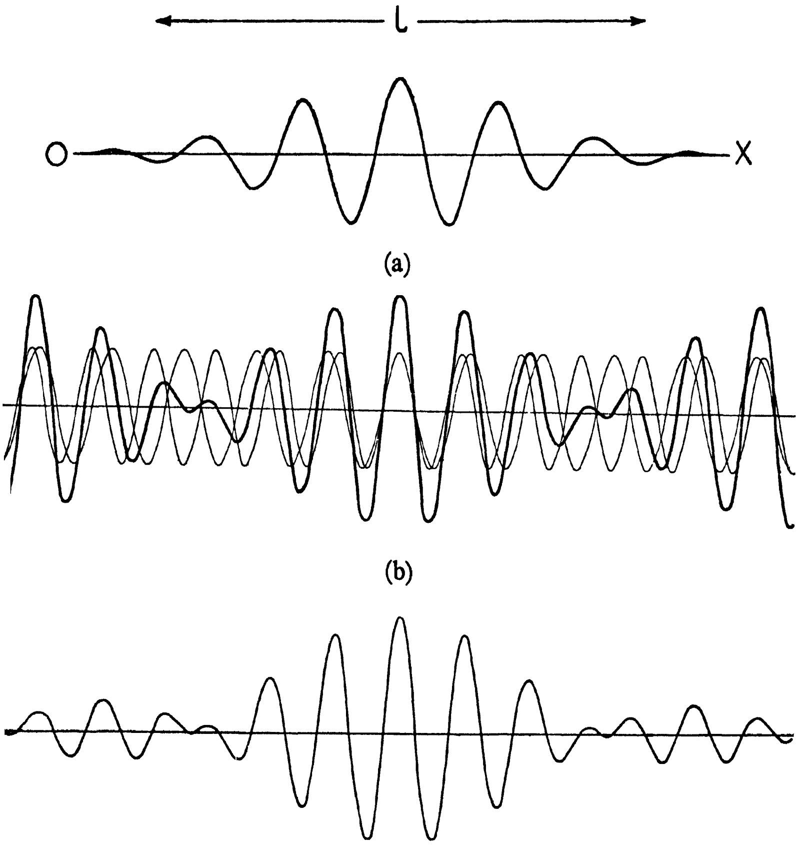


Fig. 24. (a). Wave train:
(b). Combination of two regular waves.
(c). Combination of four regular waves.

measure the field and at each point the electric field has been plotted vertically. The curve differs from that shown previously in that it does not go on oscillating indefinitely, but is limited to a fairly small region. Within this region it shows regular oscillations as in Figure 21.

A wave of this kind is known as a "wave train" or a "wave packet". Such a wave train can be built up from many extended waves of slightly different wavelengths. To see this consider first Figure 24b, in which the thin lines represent two perfectly periodic waves of slightly different wavelength, and the heavy curve has been obtained by adding the two together, so as to represent the combined effect of two such waves. This curve shows so-called "beats", i.e. it shows oscillations which are strong in certain places where the two original waves reinforce each other, and weak where they work in opposite directions. Such beats are well known, for example, in music where the combined effect of two very slightly different notes results in a warbling tone.

In this example we have by using only two periodic waves obtained a wave which still has repeated regions of strong intensity. Figure 24c shows the result of using four periodic waves, and now we find much fewer regions in which there is an appreciable intensity, because in some of the places where two of the waves reinforce each other the other two pull in the opposite direction. However, there remain still many regions of strong intensity. We can go on in this way, and by adding more and more periodic waves covering always the same overall spread of wavelength, we can spread the regions where there is any intensity further and further apart, so that we may regard the wave train of Figure 24a as an extreme case of this.

It is now easy to estimate the difference between the shortest and the longest wave that must be contained in the set of periodic waves from which a wave train like that of Figure 24 can be built. We must only remember that if $\frac{1}{2}l$ is the distance from the centre to the end of the wave train as in the figure, then the waves must be all in phase in the centre to give the strongest intensity, and must cancel each other completely a distance $\frac{1}{2}l$ away. Now if λ_1 and λ_2 are respectively the shortest and longest wavelengths we have used, then the distance $\frac{1}{2}l$ will contain $\frac{l}{2\lambda_1}$ complete waves of the one and $\frac{l}{2\lambda_2}$ of the other. Since at this point the waves must be able to cancel, these numbers must differ by at least $\frac{1}{2}$. Therefore the construction of our wave train should be possible, provided the periodic waves we are using cover a range of wavelengths such that $l\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right) = 1$. This result can be expressed more simply if we introduce in place of the wavelength the "wave number" k which equals $1/\lambda$, the reciprocal of the wavelength, or the number of waves per centimetre. The rule then becomes $l\,(k_1 - k_2) = 1$. Or in

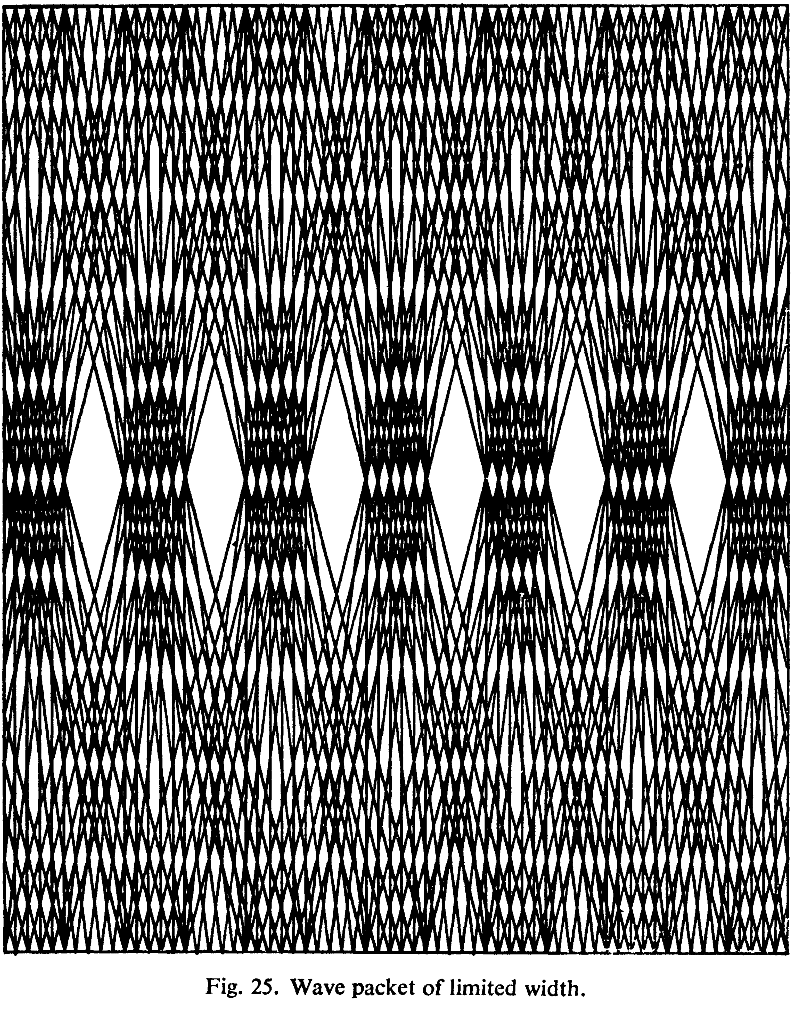


Fig. 25. Wave packet of limited width.

other words, the shorter the wave train we wish to build the greater the difference is in wave number that we have to use.

In empty space all the periodic waves of which our wave train is made travel with the same velocity, and the whole wave train therefore travels without altering its length or its nature. However, if it passes through a refracting medium, i.e. a prism, in which light of different colour, i.e. different wavelength, undergoes slightly different diffraction, the wave train will be distorted. As long as this is not the case it holds together like the "light particle" assumed by the emission theory, but when it passes through a lens system then the difference in wavelength of its constituent parts results in complications that the emission theory had not foreseen.

So far we have considered only a wave which is limited in its direction of travel, but unlimited sideways; but by an extension of the same argument we can construct a "wave packet" which also has a limited width. How this is done is explained by Figure 25 in which a number of periodic waves of slightly different directions have been indicated by shading, the shaded portions showing the wave crests with the troughs left blank. The result is a strong alternation of white and very dark regions along the middle with the rest of the picture appearing uniformly grey. This figure was constructed by using five periodic waves, and these would again lead to regions of strong intensity above and below the field shown in the picture, but by adding more and more waves with intermediate directions, all these can be cancelled out.

Here, however, there is a new feature. The waves have been drawn so that along the centre line all their crests coincide exactly. This means that the ones which are more inclined to the centre line must have a shorter wavelength. As a result the waves have different frequencies and if we consider the same wave packet after some time the different waves will have changed in appearance by different amounts and will no longer reinforce each other so neatly along the centre line. We obtain then a weaker oscillation, spread out further sideways. In other words a narrow beam of light will not indefinitely remain narrow, but will spread gradually. This spread is imperceptible if the width of the beam is very large compared to the wavelength, since then we have to use in Figure 25 only waves whose directions differ by extremely small angles.

Here again the wave theory leads to complications not foreseen by the emission picture, in this case even without the use of glass prisms or other optical instruments. For a very sharply limited beam such as at the edge of the shadow of some object seen in very parallel light, the wave theory predicts that over large distances its edge should not remain sharp, but spread out because of the spreading of the waves.

For all practical purposes, i.e. with the light sources used in practice, and objects of a reasonable size, this distortion of the shadow is quite unimportant and therefore we may think of limited wave packets of reasonably small length and width travelling as if they were particles.

The branch of optics which considers only light rays, i.e. the path travelled by such wave packets, is known as "geometrical optics". It is sufficient for use in the design of optical instruments, provided one is dealing with circumstances in which diffraction and interference are unimportant.

From our point of view these facts are interesting for two reasons. Firstly, because they illustrate the way in which progress in our knowledge about the laws of nature retains the older ideas as useful and as approximately valid in suitable situations, but shows us when we have to make essential use of the new ideas arising from our better insight.

Secondly, the relation between extended periodic waves and wave packets will later prove of importance to us when we discuss the wave nature of matter.

4

Atoms and Electrons

CHEMISTRY AND THE ATOMIC HYPOTHESIS

So far the properties of matter have entered into our discussion of the laws of nature only in a rather accidental way. We have talked in mechanics of bodies which are light and which are heavy, of substances which are soft and yield to pressure, and bodies which are stiff and strongly oppose any change of shape, of solids, liquids and gases. We have mentioned substances which conduct electricity and others which can be magnetized; substances which transmit light and of their refraction; metals which make mirrors and objects which are coloured. One could fill a chapter simply with an enumeration of the various characteristics that occur amongst the objects which we find provided by nature or which we can manufacture.

It is part of our search for the laws of nature to classify these characteristics and to find some simple principles governing them. The first part of this development belongs to the science of chemistry. Over the centuries chemists have found that all known substances are combinations of simpler ones called the chemical elements, of which 92 occur in nature.

Of the substances in everyday use not many are elements. Many metallic elements will be familiar to the reader: aluminium, copper, iron (though iron used in practice is usually very impure), tin, silver, gold, lead, etc. The gas hydrogen is an element; it is used to fill balloons and it is contained, mixed with other gases, in the gas from the mains which we use for cooking and heating. Soot is largely carbon, another element. But most of the other substances which we use are "compounds" or combinations of the chemical elements, or mixtures of them. Air is a mixture of nitrogen and oxygen; water, a compound of hydrogen and oxygen; "common" salt, a compound of sodium and chlorine, etc.

Mixtures usually behave in much the same way as their components. Nitrogen and oxygen are gases and so is air, which is a mixture of them, but if the gases hydrogen and oxygen are made to form the compound water, we obtain a liquid.

Modern chemistry started with the application of methods of measurement, above all of weighing, to the formation of these compounds.

Two rules emerge at once from this: Firstly that if known amounts of two elements, let us say hydrogen and oxygen, are turned into their compound, water, the mass of the compound which results is exactly equal to the combined masses of the elements from which we started. It is possible, for example, to combine one gramme of hydrogen with eight grammes of oxygen and the result will be exactly nine grammes of water. This rule is known as the law of conservation of mass.

The second rule is the law of constant proportions. This says that the proportion of the amounts of the elements which combine must always be the same. For example, the amount of oxygen required to combine with a given amount of hydrogen is always eight times as large. If we mix one gramme of hydrogen with more than eight grammes of oxygen and ignite the mixture, all the hydrogen will have combined with oxygen, but there will be some oxygen left over. If we start with one gramme of hydrogen and less than eight grammes of oxygen there will be some hydrogen left.

This rule of constant proportions led to the atomic hypothesis. This hypothesis, which by now has been confirmed by overwhelming evidence, says that all matter consists of atoms and that different chemical elements contain different kinds of atoms, which in particular have different weights. For most purposes we may assume all the atoms, let us say, of oxygen to be similar to each other and to have the same mass. (This statement will have to be qualified later.)

Such atoms can be connected together to form "molecules". A water molecule consists of one atom of oxygen to which two atoms of hydrogen are attached. This is indicated by the chemical "formula" for water, H_2O. From what was said about the proportions of hydrogen and oxygen in water, it is then evident that the mass of the oxygen atom must be sixteen times that of the hydrogen atom.

A molecule is therefore the smallest piece of the chemical compound, such as water, which has the chemical characteristics of water as such. It cannot be subdivided further without dividing the hydrogen from the oxygen atom and thus causing a chemical change.

Broadly speaking the branches of physics which were reviewed in the previous chapters and several others are concerned with changes which leave the molecule intact. If we subject water to forces that make it flow or pass sound waves through it, if we place it in a magnetic field or shine light on it (though not always if we pass electricity through it), if we evaporate it or freeze it, the structure of its molecules does not change. It is the province of the chemist to consider changes in which atoms are removed from a molecule or new atoms added to it, so as to form different compounds.

We now know that this division is somewhat over-simplified and that particularly in solid substances the borderline between physics and chemistry is not very distinct. Fortunately this is merely a matter of terminology, since we now know that chemistry and physics are derived from the same basic laws and are branches of the same subject, kept apart only for practical reasons.

Speculation on what would happen if we divided matter into smaller and smaller parts had been a favourite pastime of philosophers long before the atomic hypothesis assumed its important position in chemistry. There were always two schools of opinion, those who asserted that matter should be divisible indefinitely and those who expected that there would be ultimate units of matter, the atoms. The name atom itself, which means something indivisible, comes from these speculations, but while tradition has preserved the name we shall see that there is nothing at all indivisible about the atoms of modern physics.

However, the idea of atoms could not play any part in the development of our knowledge as long as we were not in a position to test its truth, to prove or disprove it. And this is possible only by linking the consequences of such an hypothesis with observations and experiment on the actual behaviour of the substances. Such observations may be purely qualitative, but they are most conclusive when they can be brought into a quantitative form and linked to accurate measurement. The development of the atomic hypothesis in chemistry is an excellent example of how exact and quantitative measurement, in this case the weighing of reacting chemicals, can show up regularities which suggest a new concept and can confirm it.

I shall not here review the reasoning by which it was decided that water contains just one oxygen atom to two hydrogen atoms, in other words that the formula for water is H_2O and not for example, HO, which would require the oxygen atom to have the same mass as 8 rather than 16 atoms of hydrogen. Broadly speaking this question is settled by the existence of compounds in which one of the two hydrogen atoms is replaced by an atom of some other element.

The use of the atomic hypothesis in chemistry proved most fruitful and convenient, and it turned out that without exception any chemical compound could be represented by a suitable formula in terms of atoms.

Yet even at this stage the existence of atoms was only a plausible and convenient hypothesis. It was not proved. The kinetic theory of heat, to which we shall return, gave further support for the atomic hypothesis. But to show that atoms were a reality it was necessary to find methods to determine the size or the mass of an atom. The chemical reasoning about the proportions in a reaction involved only the comparison between the weights of hydrogen and oxygen, and it makes no difference there whether we take the weight of both atoms (and hence of the complete molecule) to be ten times larger or ten times smaller. Towards the end of the nineteenth century, when this stage had been reached, there were still some very distinguished scientists who opposed the use of the atomic hypothesis as not well-founded and premature. These included E. Mach, who made important contributions to the foundations of mechanics and relativity.

Mach was undoubtedly right in stressing that we could not be sure of atoms until we had discovered ways of saying how big they were, but in retrospect one would say that the supporters of the theory showed more imagination in recognizing the strength of the circumstantial evidence in favour of the idea, and in expecting that the final proof would eventually be found. In any event there is no doubt whatever that if Mach and other distinguished opponents of the atomic theory were alive to-day they would at once admit that the case is proved.

THE SIZE OF AN ATOM. IONS

There are now a great many known methods of weighing atoms, of measuring their size, and counting them. Many of these methods depend on parts of physics which we have not as yet met in this book, but there is one very powerful one which can be conveniently discussed at this stage. This is the diffraction of X-rays. We have already mentioned the fact that X-rays are electromagnetic waves like light, but of a wavelength some thousand times smaller than of light. For the present we are not concerned with the method by which such X-rays are produced, but most readers will have seen an X-ray tube, whether in a laboratory, in a hospital or in a shoe shop.

It was found first by von Laue that when X-rays pass through solid matter they are strongly deflected through certain preferred angles and not at all in other directions. The pattern which results is just like the diffraction pattern from a grating, which was mentioned in the chapter on light. The obvious conclusion is that the atoms in the solid are arranged in a regular pattern, spaced equally from each other, so that the reflection of X-rays from the different atoms takes the place of the reflection of light from the different lines of the grating. This idea works out with very great accuracy, and from the observed pattern of X-ray diffraction one can deduce what the "grating", i.e. the atomic arrangement must be like. In particular the directions into which X-rays are deflected depend on a comparison between the distances of adjacent atoms from each other and the wave lengths of the X-rays. By means of X-ray studies we can therefore measure the atomic distances in a solid in terms of X-ray wavelengths.

This does not solve our problem unless we can measure X-ray wavelengths. This is possible by measuring the diffraction of X-rays on a grating which is made by ruling fine lines on a metal surface. Of course the finest lines we can make in this way will be much wider than X-ray wavelengths but we can still obtain an interference pattern by using small angles in just the same way in which we had seen the diffraction of light from a gramophone record at small angles.

This is the principle of one of the accurate determinations of inter-atomic distances. By this method we can find out how many atoms there are, for example, in a cubic centimetre of some solid, let us say common salt. And since we also know the weight of a cubic centimetre of salt we find in this way the weight of one atom.

The distances between atoms in a solid come out to be a few times 10^{-8} cm., in other words a few thousand times smaller than the wavelength of visible light. It is therefore clear that we shall never be able to see atoms directly, however powerful the microscopes we may build, since we get clear images of objects through a microscope only when the outline of the object can be traced by the methods of geometrical optics, which were discussed at the end of the last chapter. This requires that the object be large in comparison with the wavelength of the light, whereas the opposite is true of atoms.

The weight of an atom of hydrogen turns out to be $1{\cdot}6 \times 10^{-24}$ gms., a number which is so small that it is very hard to visualize. Perhaps the following consideration may give an impression of how small atoms are: if we pump out air from a space by means of a powerful pump we usually think of such a "hard vacuum" as very empty if we have reduced the amount of air in the space 10^9 (i.e. a thousand million) times. Even then, however, every cubic millimetre still contains about twenty million atoms.

Some of the other methods of counting atoms or measuring their dimensions will come up later on, but the main thing is that all these methods give the same answer and thus confirm that an atom is not just a convenient fiction, but a real measurable and weighable object.

To learn more about atoms one has to study their internal structure. The beginnings of this came from work on the passage of electricity through liquids and gases. Faraday had studied the passage of electric currents through certain liquids, in particular through water in which some kind of salt was dissolved, and it was found that as an electric current passes through such solutions from one metal plate to another some of the chemical constituents of the salt would accumulate at the plate which was kept at a positive potential, and others on the negative plate.

This process, which is known as electrolysis, is used in electroplating and is also closely related to what happens when the accumulator or "battery" of a motor-car is charged.

It was further found that the quantity of each substance deposited in this way is directly dependent on the amount of electric charge which has been passed through the liquid.

In the light of atomic theory these regularities are interpreted as a tendency of each of the atoms in the salt to acquire a definite electric charge, positive or negative. For example, common salt is a compound of sodium (chemical symbol Na) and chlorine (symbol Cl) containing equal numbers of atoms of each with the chemical formula NaCl. In a solution of this in water some of the NaCl molecules are split into sodium and chlorine in such a way that the sodium carries a fixed amount of positive electric charge and chlorine the same amount of negative charge. Such charged atoms are called "ions". If an electric field is applied through metal plates in the solution the sodium ions are attracted by the negative plate and the chlorine ions by the positive plate. On arrival at the plates they are neutralized, i.e. they lose their charge. Neutral atoms of chlorine or sodium will remain in some combination with water.

Clearly if each atom of sodium always carries the same amount of electric charge, the total charge that has passed from one plate to the other will be in a definite relation to the amount of sodium that has been removed from the solution.

In this way, we can find how much charge corresponds to one gramme of sodium. Since we now know the weight of the sodium atom we can deduce how much charge a sodium ion carries in the passage of electricity through the solution. The result is that the charge carried by a sodium ion and that of a chlorine ion are opposite and equal, and that the same amount occurs in all other salts and all other such electrolytic processes.

Some ions carry two or three times the basic charge, or even a greater multiple.

These conclusions suggest that not only matter, but also electricity has an atomic structure and there exists a smallest electric charge.

Similar conclusions follow when we study the conduction of electricity through gases, as in the discharge tubes which are now familiar in advertising signs or in fluorescent lighting. Here again we find ions carrying electric charge. The case of a low-pressure discharge tube has the advantage that the mechanical problem of the motion of such ions is now simple. In a liquid an ion is always very close to other atoms and molecules, and its passage is much like that of a person in a jostling crowd. In a low-pressure discharge tube the ion may travel a long way without encountering any other atoms or ions. It is therefore possible to apply, for example, electric forces to these ions and observe how much they are deflected from their path. If we knew the speed of such an ion the deflection would tell us how much it had been accelerated sideways and therefore would give us the force divided by the mass, or since we know the field strength, the quantity $\frac{e}{m}$. i.e. the electric charge divided by the mass.

In practice we do not know the speed of the ions, but we can find it by applying also a magnetic field, for instance by placing the discharge tube between the poles of a magnet of known strength. A moving electric charge represents an electric current and we have seen in Chapter 2 that a magnetic field exerts a force on current which is stronger the stronger the current. It follows therefore that charged particles in a magnetic field experience a force at right angles to the direction of motion whose strength increases with the velocity of the particle.

THE ELECTRON

This is a rough sketch of the idea behind J. J. Thomson's method of measuring $\frac{e}{m}$ for various ions in a gas discharge. It was found that there were many ions with positive or negative charges for which the relation between charge and mass was just the same as that found in the observations on electrolysis. But in addition, and in fact most prominently, there occurred rays that appeared to consist of particles which for the same electric charge were nearly two thousand times ligher than the lightest atom, i.e. that of hydrogen. Their electric charge was always negative. These particles were called "electrons".

The rays consisting of fast electrons in a discharge tube are known as cathode rays, because they come from the cathode (which is the name for the metal part of the tube which is kept at a negative potential).

Further work on the passage of electricity through gases showed that if a gas at first contains only atoms which are electrically neutral, i.e. carry no electric charge, it is possible to produce in it electrons and positive ions. This showed that the electrons were contained in these atoms; an atom which has lost a negative electron then carries a net positive charge, and becomes a positive ion. Negative ions are similarly formed when an electron attaches itself to a neutral atom. Ions therefore always carry a charge which is either equal in amount to the charge of an electron (but either positive or negative), or they can carry a multiple charge by losing or acquiring two or more electrons.

It is comparatively easy to measure the electron charge, which is usually denoted by the letter *e*. The most elegant way of doing this is due to Millikan. He looked through a horizontal microscope at a fine mist of oil, consisting of microscopic oil droplets. Such droplets fall under their own weight, but owing to their small size they fall extremely slowly. This is because the weight of such a drop is proportional to its volume, i.e. to the third power of its diameter, whereas the resistance of the air to its motion varies as its cross-section, i.e. as the square of its diameter. For a very small drop therefore the air resistance is much more important in relation to the weight than for a large drop. In Millikan's apparatus the oil drops were between two horizontal plates which could be electrically charged, as in a condenser. One then finds that some of the drops carry an electric charge and therefore would fall more rapidly or less rapidly, or even rise, as the electric field is applied, according to the magnitude and sign of their charge. By watching through the microscope the motion of some particular drop with and without electric field, it was therefore possible to deduce the electric force on it, and as the electric field strength was known, the magnitude of the charge.

This charge was then found to be always an exact multiple of a smallest charge. This is interpreted as the charge of an electron, since the easiest way for the drop to become charged is by gaining or losing an extra electron. The charge of the electron turns out to be extremely small, about 5×10^{-10} in the units used in Chapter 2. In language which is likely to be more familiar this means that current of one ampere represents the passage of 6×10^{18} electrons per second.

Knowledge of the electron charge gives us another determination of atomic masses. We had seen that in the passage of electricity through liquids the relation of the charge that has passed to the quantity of material deposited was constant and depended on $\frac{e}{m}$ where e is the charge of an ion and m its mass. Since the simplest ions carry just one electronic charge the knowledge of this charge gives us their mass.

The discovery of the electron at the end of the nineteenth century was the starting-point of the study of the structure of the atom, since the electron was evidently one of its constituents.

WHAT ARE ATOMS MADE OF?

It took quite some time to find out more about the interior of the atom and the precise part played in it by the electron. The next advance came from the work of Rutherford. He was studying radioactivity, and to appreciate the new evidence about the atom which he obtained we must say a few words about this phenomenon. It had been known from the work of Becquerel that there exist minerals, in particular the so-called pitchblende, which send out very penetrating rays. These can blacken photographic plates, they can make air conduct electricity and they can be detected in several other ways. Marie and Pierre Curie isolated from pitchblende the element radium, which forms a particularly strong source of such radiations. They also showed the existence of several other radioactive substances.

Rutherford studied in more detail the nature of the radiations which come from these substances. They could be divided into three main types, which Rutherford denoted by the first three letters of the Greek alphabet as alpha, beta and gamma-rays. He showed that the alpha rays could be deflected by an electric or magnetic field in the same way as beams of positively charged particles, the beta rays behaved like negatively charged particles and the gamma-rays were not affected by electric or magnetic fields at all.

It was found that gamma-rays consist of electromagnetic waves of a wavelength much shorter even than that of X-rays. Beta-rays consist of electrons, and the alpha-rays consist of positively charged heavy particles. Measurement of their deflection in the field showed in fact that their charge was in the same proportion to their mass as for ions of the gas, helium, which is the lightest element next to hydrogen. We know to-day that alpha particles actually are ions of helium.

These are thrown out by radium and similar substances with very high speeds, sometimes as great as $\frac{1}{10}$ of the velocity of light, and they are able to pass through thin sheets of solid substances, for example a sheet of paper or aluminium foil. We shall return later to the question of the origin of these rays.

Rutherford asked to what extent these alpha-rays would be deflected from a straight path by their passage through matter. He therefore directed a fine beam of such particles on to a thin foil. The result was, somewhat unexpectedly, that most of these particles had not been noticeably deflected. They continued in a straight line. Only a few showed noticeable deviations from their original direction and even fewer were deflected through large angles.

This result was most surprising. We know that solid matter must be closely filled with atoms since even very high pressures will not reduce its dimensions appreciably. The alpha particles of Rutherford therefore must have passed right through a large number of atoms and one would have expected them to collide with parts of these atoms. In such collisions, they would be deflected from their straight path. Certainly a billiards player would be most surprised if he succeeded on a table filled completely with billiard balls to project another ball from one side and have it emerge directly in a straight line on the other side. This in fact would be possible only if there was plenty of room between the different balls for the cue ball to pass through.

Rutherford thus concluded that the atom was mostly empty and that all heavy parts which it contained, as well as the alpha particles which he was using as his cue ball, must be very small compared to the size of an atom.

One could say more than that. The few alpha particles which had been deflected evidently had happened to pass close to some part of the atom where relatively strong forces had acted on them. Since the alpha particles themselves were electrically charged, it was natural to think of electric forces as being responsible for this deflection.

By a brilliantly simple analysis Rutherford showed that the numbers of alpha-particles deflected through various angles corresponded with what one would expect if each atom contained a very concentrated electric charge whose force on the alpha particle would follow the inverse square law as in Chapter 2. This concentrated electric charge had to be attached to a mass heavier than that of the alpha particle, since otherwise the alpha-particle would just have pushed it out of the way without being greatly deflected itself.

These arguments suggested the following picture of the atom: The atom contains in its centre a small positively charged core or "nucleus", which is surrounded by electrons. The number of electrons is such that the atom as a whole is electrically neutral. In this case the electrons would move under an attractive force following the inverse square law, like planets in the attractive field of the sun, and the model of the atom which Rutherford proposed made it look like a solar system in miniature. The atom is in fact just about as empty as the solar system, i.e the atomic nucleus is as small in comparison to the size of the atom as the sun in comparison to the orbits of the planets.

The mass of a nucleus is therefore nearly the mass of the whole atom; even for the lightest atom it is nearly two thousand times the mass of an electron.

If the atom is usually to have no net electric charge, the electric charge of the nucleus must just balance that of all the electrons. If the nucleus has a charge of $+Ze$ where e is again the electronic charge and Z is a whole number, it must have Z negative electrons to balance it.

This number Z linked up at once with some surprising regularities found in chemistry. Mendeleyev had found that if the known chemical elements are written down in the order of their atomic weights (as determined from their proportions in chemical reactions) there appeared a periodic repetition of certain typical chemical characteristics.

For example, the list will contain in different places the "noble" or "inert" gases helium, neon, argon, etc., which remain gaseous down to fairly low temperatures and do not form any compounds with other elements at all. Now the element following on such a noble gas is always an alkali, i.e. helium is followed by lithium, neon is followed by sodium, argon by potassium and so on. All these are metals which are chemically most active and react even with water, they form salts, and in solution they easily form positive ions, i.e. easily lose an electron.

On the other hand, the elements preceding the inert gases (with the exception of the first) are all of the chemical type known as halogens, they are fluorine, chlorine, bromine, etc., elements which in combination with hydrogen form acids, and in solution form negative ions, i.e. easily pick up an additional electron. These regularities led Mendeleyev to construct the "periodic system" of elements.

It now turns out that the number Z which is the charge on the nucleus in units of the electron charge is also equal to the serial number of the element in Mendeleyev's table.

The lightest element, hydrogen with $Z = 1$, therefore contains a positive nucleus and one electron of equal negative charge. The atom of the second element, helium, has twice as much electric charge on its nucleus, and when the atom is neutral it contains two electrons. This list continues up to the heaviest element known in nature, uranium, with $Z = 92$, i.e. a nucleus containing 92 times as much positive electric charge as that of hydrogen and surrounded by 92 electrons.

WHY DON'T ATOMS COLLAPSE?

On the basis of this picture, we now try to understand the behaviour of an atom. Take the simplest case, that of the hydrogen atom. It contains the hydrogen nucleus, or "proton", and one electron. The two attract each other according to the inverse square law, and the mechanical problem is therefore the same as for the sun and one planet. We would therefore expect that the electron will revolve around the proton. As long as there was no disturbing influence present this would be a perfectly stable state of affairs and the shape and size of the curve described by the electron, its "orbit", would remain unchanged, just as the orbit of a planet about the sun remains the same as long as there are no disturbances.

However, in the case of an atom we have every reason to believe that there should be disturbances. Two influences are particularly important. The first is the influence of other atoms, particularly in solids or liquids in which the atoms are packed very closely together. The forces between adjacent atoms then represent a strong disturbance; each electron should no longer follow the ideal steady orbit, and the result should be to make all the electrons gradually approach their respective protons closer and closer, so that through the attractive force more and more energy would be released; at the same time the size of the electron orbit and with it the size of the atom would diminish, leading to a complete collapse of the solid or liquid substance. This, of course, is quite unlike the behaviour of real substances and therefore there is something seriously wrong with our picture.

Even if we think only about a single atom, as we may do in the case of a gas, in which the atoms are most of the time fairly far from each other, there would still remain another cause of disturbance to the simple motion of the electrons. This is the emission of light or other electromagnetic waves. An electron, i.e. a charged particle, revolving in an orbit produces an oscillating electric field which leads to the emission of electromagnetic waves, just as in the case of the simple aerial which we discussed in Chapter 2 (Figure 17). The radiation which is caused by this carries away energy. By the law of conservation of energy it follows that the mechanical energy of the electron must diminish all the time. This means that the orbit of the electron should gradually shrink and this should continue until the electron has "fallen" into the nucleus. This argument again demonstrates that the laws of nature as we know them up to this point cannot explain the existence of atoms of definite size and with properties which do not change with time.

These two arguments are examples of the difficulties which were met in the attempt to explain the structure of the atom in terms of the laws of "classical" physics, i.e. the mechanical laws of Newton and the electromagnetic laws of Maxwell. In a later chapter we shall see that this and other evidence calls for a very profound revision of the fundamental laws, involving not so much a change as a refinement, which is the subject of the quantum theory.

MANY FACTS CAN BE UNDERSTOOD

However, in spite of these contradictions even at the present stage of incomplete knowledge which we have reached, a good many facts about the behaviour of matter can be understood in terms of atoms. For example, if atoms contain, as we have seen, positive and negative electricity then if an atom is placed in an electric field this will pull on the positive charges one way and on the negative charges the other way and therefore will shift the one with respect to the other. In terms of electrons revolving in orbits about a positive nucleus, we can say that the orbits will be shifted so that the electrons are no longer distributed evenly around the nucleus. Now if we are not dealing with a single atom, but with a solid piece of matter containing very many atoms, the same will apply to each atom. The net result is that all positive charge in this piece of matter gets displaced a small distance one way and negative charge the same amount in the opposite direction. Now since in the interior of the piece of matter any small volume contains equal amounts of positive and negative charge their displacement does not alter the charge density there, but it will make a difference near the surface; in the result we end up with a layer of positive charge on the one face and a layer of negative charge on the other. Figure 26a indicates a slab of material containing positive and negative electric charges. In 26b this has been placed between two metal plates, charged positively and negatively, as in the "condenser" discussed in connection with Figure 11. All the positive charges are moved up, all the negative charges down and the net result is an excess of positive charge at one end and an excess of negative charge on the other. These layers of "induced" charge produce an electric field which counteracts that produced by the plates. As a result the field between the plates of a condenser is diminished if the gap between the plates is filled with a suitable non-conducting material. This means in turn that the potential difference for given charge is reduced, or that at the same potential difference the condenser contains more charge.

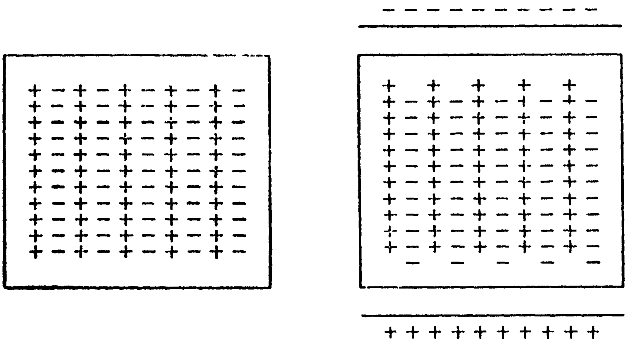

(a) (b)

Fig. 26 (a). Matter contains positive and negative charges.
(b). Displacement of charges by an external field.

We say that its "capacity" is increased. The tendency of a material to counteract in this way an external applied field is connected with its "dielectric constant". In formulating in Chapter 2 the laws of the electromagnetic field I have made no reference to the dielectric constant although it is important in practice for dealing with the fields inside or near dense substances. The reason was just because this is not part of the fundamental laws, but reflects the structure of these substances and the effect of the field on the atoms contained in them.

In the same way one can understand that magnetic fields must influence the motion of the electrons, though the precise way in which the orbits are affected is not simple.

Also the fact that some materials can conduct electricity is understandable, provided we can believe that in such substances electrons can get separated from some or all of the atoms and move freely inside the body. This is supported by the fact that conductors of electricity are usually elements which the chemists know as electropositive, i.e. which easily form positive ions by the removal of one or more electrons. It has also been shown directly that the electric current in a metal is carried by the movement of negative rather than positive charges and that their mass corresponds to that of electrons rather than that of atoms.

Apart from these electric and magnetic problems, the concept of atoms is capable of throwing light also on the structure and the ordinary mechanical properties of matter, but since these depend on the temperature we cannot say much about them until we have outlined the main ideas of the connection between heat and the motion of atoms.

5

Atoms in Crowds

HEAT AS IRREGULAR MOTION

QUANTITY OF HEAT. HEAT AND DISORDER

THIS chapter will be about heat, a branch of physics which, unlike electricity, concerns changes directly accessible to our senses and familiar from everyday life.

Two basic concepts in the study of heat are the temperature and the quantity of heat. Every reader will know what is meant by temperature, since he must have had occasion to read a thermometer and to experience the different feelings of discomfort that are caused by either too high or too low a temperature. The idea of the quantity of heat is less obvious.

When some object, let us say a pan full of water, is heated, i.e. raised to a higher temperature, we say that it contains a greater quantity of heat. However, different bodies brought to the same temperature contain different amounts of heat. First of all if two kettles contain different amounts of water then the fuller one will require a greater amount of heat to raise it to the boiling point than the one containing less water. Heat is measured in terms of calories. A calorie is the amount of heat necessary to raise the temperature of one gramme of water by one degree centigrade. Alternatively one British Thermal Unit, or B.T.U. is the amount of heat that is necessary to heat one pound of water by one degree Fahrenheit.

However, if instead of one gramme of water we have to heat one gramme of something else, say of lead, by one degree, a different (in this case smaller) amount of heat is required. This is shown by the following kind of observation: We take two vessels containing the same quantity of cold water. Into one we pour a certain quantity, say one kilogramme, of water heated to boiling point and stir the mixture so that it reaches a uniform temperature. Into the other we immerse one kilogramme of lead which has been heated to the temperature of boiling water and wait until water and lead come to the same temperature. We measure the temperature of both vessels after this and find that the first is warmer than the second. This means that the water brought with it a greater amount of heat than the lead. We say then that lead has a smaller "specific heat" than water.

Having thus defined the quantity of heat, we may link this with the fact, noted in Chapter 1, that heat is a form of energy. If we stop the motion of a motor-car by applying brakes, the brake drums become warm. If we measure this rise of temperature in circumstances when the cooling of the brakes by the air and by their contact with other parts of the car is unimportant, and allow for the amount of metal contained in the brake drums and for its specific heat, we can find the quantity of heat produced. This we compare with the amount of kinetic energy which the car has lost and it turns out that the heat generated is exactly proportional to the amount of mechanical energy lost. It is reasonable to say therefore that mechanical energy is being transformed into heat.

We have already stressed that it is possible to transform mechanical, electrical, magnetic, chemical and heat energy in various ways into each other. But there is something peculiar about heat, and that is the "irreversible" character of many processes involving heat. For instance it is easy to convert mechanical energy into heat, as in friction, but the reverse is not so simple. No amount of heat applied to the brake drums of a car will set the car into motion again. It is easy to convert electricity into heat by letting a current pass through a resistance wire, but it is very hard to convert heat directly into electricity; heating a wire will not start a current; though we know of thermo-couples, combinations of different metals which do produce an electric potential difference when heated, this is a very small effect. It is easy to convert chemical energy into heat as when we are burning coal, i.e. combine carbon and oxygen to carbon dioxide, but difficult to break up the carbon dioxide back into carbon and oxygen.

This is all quite different from the conversion of electricity into mechanical energy by means of a motor, since we saw that an electric motor may also be used as a generator. It is different from the conversion of chemical energy into electrical in an accumulator, since by recharging the accumulator we convert electrical energy into chemical.

We meet the same kind of irreversibility in the process of heat conduction. When two objects, one hot and one cold, are placed in contact with each other, heat will flow from the hotter to the colder until their temperatures are equal. There are good and bad conductors of heat, which differ merely by the speed with which the temperatures will equalize. Amongst the best conductors of heat are metals (that is the reason why metals feel colder to the touch, since they conduct more heat away from our hand). Here again heat will flow from the hotter to the colder body, but never the other way, so that the process cannot be reversed; once the temperatures of two bodies in contact have become the same we cannot make them different again unless we apply heat from outside to one of the bodies or take some away from the other.

For this reason, the knowledge of the total energy which a body or a piece of machinery contains, including its heat energy, is not enough to tell us how useful this energy can be in practice. For instance, if I have a large block of matter which is very hot, say above the boiling point of water, and another which is cold, then this is a useful source of energy, because I can use the heat by pouring water on the hot object to raise steam. This can drive a steam engine, and the steam that comes out of the engine can be condensed to water again and returned for further use. On the other hand if I allow the hot and cold block to equalize their temperatures by heat conduction they will then still contain the same amount of energy, but this will be useless. We normally tend to think that heat, e.g. of a boiler, is enough for raising steam and working an engine, but we take it for granted that besides this we also have a source which supplies water at ordinary temperature; this temperature difference is essential. If we lived on the surface of a planet which was so hot that all water were evaporated this would not allow us to turn this heat into mechanical energy by means of a steam engine or in any other way.

One therefore introduces, besides the energy, another quantity which is called the "free energy". This is a measure of that part of the total energy of an object, or a number of objects, which can be turned into other useful forms of energy. Mechanical, electrical and magnetic energies always count in full towards the free energy, but only part of the heat energy is to be included. In our example of the hot and the cold block the free energy is obviously larger when they are at different temperatures, than when their temperatures have become equal. As heat is transferred by conduction from the hotter to the colder body the hot body loses as much heat as the cold one receives, but since the free energy is reduced, the hot body must lose more free energy than the colder body receives. The difference between the total energy and the free energy is therefore a measure of how much heat is stored in a practically useless form. This quantity, the difference between the total and the free energy is closely connected with what is called the "entropy". It is not necessary for our purpose to explain precisely what is meant by entropy or how it is measured, but it is important only to stress its close connection with the difference between actual and free energy.

If we lose free energy without making it serve a useful purpose, as in the case of heat conduction, or in the process of stopping a moving object by friction, the difference between the two kinds of energy, and with it the entropy, increases. In fact it is true that in any change that takes place in a group of objects which are undisturbed by any influence from outside the entropy may only increase, but never decrease. A change in which the entropy does not alter is a reversible change and we can usually arrange it to proceed in the opposite direction. Any change in which the entropy has increased cannot be reversed.

This statement about the increasing entropy summarizes the so-called second law of thermodynamics, i.e. of the science of heat. The first law is simply that of conservation of energy, including heat, which we had already met.

Entropy is very closely similar to disorder. The way in which it is easy to let the temperature of two bodies become the same, but hard to make them different, is very similar to the way in which it is easy to let order turn into disorder, but difficult to reverse the process. For example, it is easy to shuffle a pack of cards, but it takes much more trouble to rearrange the cards in their proper sequence. It is easy to convert an orderly state of affairs, in which we have sugar in one jar and salt in another, into a disordered state in which the two have been mixed together. It is practically impossible to go back from this disordered state to the ordered one. The impossibility of unscrambling an omelette is a familiar metaphor. In shuffling a pack of cards or mixing sugar and salt, or scrambling an egg, we are increasing the entropy. We shall see later that entropy and disorder are not merely similar, but that they are one and the same thing.

HEAT AND ATOMIC MOTION

The question now arises whether heat, like electricity, is something basic that we have to include in our formulation of the basic laws of nature, or whether it can be explained in terms of other known things. It turns out that it can be completely explained in terms of atoms and their motion. This is accomplished by what is usually called the kinetic theory of heat, or statistical mechanics, which was developed during the 19th century.

We have seen already in the preceding chapter that matter consists of atoms, which may be separate from one another or tied together into molecules. Most common gases consist of molecules. For example, air contains molecules of oxygen, each containing two oxygen atoms and nitrogen, each containing two nitrogen atoms. What is then the difference between cold and hot air? According to the kinetic theory the difference is that in hot air the molecules are moving about more rapidly.

If we could look at a small space filled with air through a much more powerful microscope than any we possess in practice, we would see this space filled with molecules darting about in an irregular manner, like a swarm of insects that has been disturbed. Whereas, however, the insects can avoid colliding with each other, each molecule travels by the laws of mechanics in a straight line with uniform speed, until it collides, either with another molecule, or with a wall of the space which encloses the air. In air at ordinary pressure there is a fair amount of room between different molecules. Since by the methods of the previous chapter we can estimate the size and the weight of a single molecule, we can find out that in air the distance of any molecule from its nearest neighbour is about ten times its diameter, and it can usually travel for something like a thousand times its diameter before it makes a collision. In such a collision, which we have to regard as elastic, in the sense of the discussion in Chapter 1, the two colliding molecules both change their direction and in general also their velocities, but in such a way that their total energy remains the same.

The path of any one molecule is therefore exceedingly complicated. It travels in a straight line only for very short distances, of the order of a few ten-thousandths of a centimetre on the average. It then collides with another, alters its velocity and its direction in some way, goes on until it makes the next collision, and so on. In principle, if we could, through our imaginary microscope, determine the exact position and the exact motion of all molecules, we could work out which of them are going to collide and where and what their motion after the collision will be. Remembering the enormous number of molecules that in fact we are dealing with, this would be a hopeless undertaking in practice, even if we possessed sufficiently fine instruments to make the necessary observations. But just because the number is so large we are not interested in the fate of a single molecule. For explaining the physical behaviour of air we need to know only what happens to large numbers of molecules on the average.

For example one of the important quantities in the behaviour of a gas such as air is the pressure. Every gas exerts a pressure on the walls that contain it and prevent its further expansion. In ordinary air the pressure on a wall is not noticed because there is usually air on the other side of the wall as well, which exerts exactly the same pressure in the opposite direction. However, if the space on one side of the wall is empty, i.e. if the air has been removed by a vacuum pump, then there is a one-sided force which is capable of breaking the wall, if it is not strong enough. As is well known this force in ordinary air amounts to the weight of one kilogramme for every square centimetre of wall surface.

In the molecular picture this pressure is due to the fact that the wall is being continually bombarded by molecules. These molecules rebound after colliding with the wall. This means they change their velocity. From Newton's laws there must therefore have been a force on them when they came in contact with the wall, and, by the law of action and reaction, an equal and opposite force on the wall. This force can be calculated if we know how many molecules are striking the wall per unit time and what their speed is. If we increase the speed of the molecules, the force exerted by each of them in the collision is increased and also the number of collisions is increased, because they will hit the wall more frequently. Therefore the pressure is proportional to the square of the velocity of the molecules. It also depends on their mass; heavier molecules will need a greater force. Therefore the pressure is proportional to the mass times the square of the velocity, in other words to the kinetic energy of the molecule. Now it is known that for a given amount of gas the pressure increases with the temperature. This is the same thing as saying that the tendency of a gas to expand increases with the temperature. Indeed if we close an "empty" bottle, i.e. a bottle containing air, with a cork and heat it, there will come a point when the pressure of the air on the cork exceeds that of the air outside sufficiently to force the cork out of the bottle. It follows therefore that the molecules move with greater speed at the higher temperature.

This statement can be made more quantitative. According to the well-known law of Gay-Lussac, the pressure of a given amount of air in a given volume at a temperature of t degrees centigrade is $p = p_0\left(1 + \frac{t}{T_0}\right)$ where p_0 is the pressure of the same amount of gas at zero degrees centigrade and T_0 is a constant which amounts to approximately 273 degrees. If we therefore cooled the air to T_0 degrees below the freezing point, so that t would equal -273, then the gas would exert no pressure at all. Actually if we tried to cool ordinary air that far it would liquify long before this point has been reached, but by starting not from ordinary air, but from air which has been reduced in amount by a vacuum pump, i.e. which has a smaller p_0 we can reach a lower temperature without it liquifying. In principle therefore we can approach the state of no pressure very closely. The temperature $-T_0$ is then, according to our previous reasoning, that temperature at which all the molecules would stop moving. This point is usually called the absolute zero of temperature. It is convenient to use a new temperature scale, called the "absolute" temperature, which uses the same divisions as the centigrade scale, but counts from the absolute zero, instead of the freezing point of water. The absolute temperature $T = t + T_0$, if t is counted from the usual zero.

In terms of absolute temperature the law of Gay-Lussac then reads $p = p_0 \cdot \frac{T}{T_0}$, or the pressure is proportional to absolute temperature. Since we had seen that for a given quantity of gas the pressure is proportional to the kinetic energy of the molecules, we therefore conclude that the kinetic energy of the molecules in a gas is proportional to the absolute temperature.

So far our reasoning provides no possibility of either proving or disproving the kinetic theory, since we have simply used the observed relation between temperature and pressure to relate the temperature to the speed with which molecules move, but the first test of the hypothesis arises when we compare different gases. It is known that for the same pressure the mass of gas contained in a given volume is proportional to the molecular weight of the gas. For example, at the same pressure a litre of oxygen weighs 16 times as much as a litre of hydrogen, and the mass of the oxygen molecule consisting of two atoms, is sixteen times that of the hydrogen molecule which consists of two hydrogen atoms. This means that for the same pressure and temperature, different gases contain exactly the same number of molecules in a given volume. Since pressure depends, as we have seen, only on the number of molecules times their kinetic energy, it therefore follows that at the same temperature the molecules of different gases also must have the same kinetic energy.

What does our theory say about this? To answer this we have to consider what happens if in the same container there are molecules of different kinds, in particular of different mass. If they all had the same velocity, this state of affairs would not persist. For example, in a head-on collision of a heavy and a light molecule with the same speed, the heavier one will change its velocity only very little, whereas the lighter one will rebound and thereby increase its velocity considerably. We have to generalize this result by considering not merely head-on collisions, but collisions at all angles. When this is done it turns out that in most of these collisions the lighter atoms increase, the heavier atoms decrease, their velocity and that this goes on until a steady state has been reached, in which on the average both kinds of molecules have the same kinetic energy. In this respect therefore the kinetic theory is in agreement with the observed gas laws.

We find therefore that according to observation and equally according to the kinetic theory the average kinetic energy of any kind of molecule is a constant times the absolute temperature and it is usually written in the form $\frac{3}{2}\, kT$ where k is known as Boltzman's constant. The reason for the factor $\frac{3}{2}$ in this definition of Boltzman's constant is only that this simplifies other formulae in which the same constant appears.

CERTAINTY AND THE LAWS OF CHANCE

It should not be imagined, of course, that all the molecules in the gas have the same energy. We have already stressed before that the path of a single molecule is a most complicated and irregular thing. Nevertheless the kinetic theory allows us to make firm predictions about how a large crowd of molecules will behave. In the example so far discussed we can talk with confidence about the mean kinetic energy of a gas of molecules and the pressure exerted by them. The reason for this lies in the law of averages. In studying the behaviour of a gas we are in a similar position to the management of a casino in which roulette is played. The motion of the ball over the roulette wheel is too complicated and irregular for anyone to foresee where it will land. But in the long run one can be sure that each number will, on the average, occur equally often. The management, whose plans depend on this being true, can count in the long run on a steady profit. Probably a series of thirty or forty successive "blacks" could break the bank if there happened to be a player who gambled on black and let his winnings stand over each time as the new stake, but the likelihood of such a series happening is so small that it can be completely ignored. The risk of the Casino in Monte Carlo going bankrupt because of such an exceptional run is less serious even than the risk of it being wiped out by an earthquake, or some other catastrophe of nature.

We can use the laws of chance for making sure predictions whenever we are dealing with large numbers. In our last example the large number is that of the run of the roulette wheel over the years. Another familiar example is that of life insurance. The survival of an insurance company depends on their knowledge of how many of their policyholders will die in the next few years. While we cannot forecast the death of any particular person, we know that in dealing with large numbers the percentage of people who will die in a given period is well predictable (except for wars, epidemics and similar events which change the expectation of life of everybody). In describing this situation it is hard to improve on Sir Arthur Eddington's famous phrase "Human life is proverbially uncertain; few things are more certain than the solvency of a life-insurance company".

It is characteristic of the kinetic theory of heat that all its statements are of this statistical kind, but this does not make them any less certain. Molecules are far more numerous than the number of clients of even the biggest insurance company, and they are not liable to unforeseen disasters like wars and epidemics, since their behaviour is controlled not by man-made institutions, but by the laws of nature.

It is for example perfectly compatible with the laws of mechanics that at a given moment there might be no molecules hitting the walls of a vessel at all, because they might all happen to be in a different part of the vessel, so that the pressure on the wall would disappear. Or else it might happen that twice the normal number of molecules hit the wall in any given time, or that those which do hit it happen to have twice their normal velocity, and such an excess of pressure could break the vessel. But in practice we need not worry about this happening, because the likelihood is enormously smaller even than that of the same number turning up 100 times in succession in a roulette game, or of all the policyholders of an insurance company dying on the same day.

In fact, the laws of probability say that the uncertainty with which we can foresee the number of times some event will take place goes with the square root of that number itself. For example if from the birth-rate in a certain town we work out that on the average 100 children are born there every week we must attach to this statement an uncertainty of about the square root of 100, i.e. 10. We should therefore expect that the number of births in that town in many weeks will be 90 or 110, and even to find 120 or 80 in any week will not surprise us, but it is most unlikely that 200 or 50 births will occur in any one week.

To return to the molecules in air. The number that at any given instant are just hitting one square centimetre of wall surface, i.e. are at that instant in actual mechanical contact with the wall, is something like 10^{12} or one million million. The square root law applies again and therefore at different times this number differs by an amount of the order of its square root, which is 10^6 or one million. In other words, the pressure on one square centimetre of wall surface exposed to air varies by about one part in a million. There are in fact no instruments with which such a small change in the instantaneous pressure could be detected.

However, there exists a way of making these changes apparent and this is by looking at the pressure not on a square centimetre of wall surface, but on a very small dust particle, which is just observable under a powerful microscope. It has long been known that such small particles move about irregularly and this phenomenon has been given the name of Brownian motion. Suppose we think of a dust particle of one thousandth of a millimetre in size, so that its surface area is 10^{-8} sq. centimetres. In that case the number of air molecules hitting it at any given instant from one side is something like ten thousand. By the square root rule we expect this number to be different at different times by about 100 or by one per cent. For the same reason the pressure of the air on different sides of the dust particle may differ by as much as one per cent; sometimes the force from one side will be greater and sometimes from the other. The result will clearly be an irregular motion of the dust particle, and this is just what is observed. We may compare such a small particle with the business of a small insurance company with very few policy-holders; their profits will vary a good deal from one year to another, because the numbers of new clients in different years and also the numbers of deaths are bound to vary a good deal.

Kinetic theory can be applied to treat the problem of Brownian motion and the results are again confirmed by observation. A particularly simple way of looking at this phenomenon is to say that the dust particle may itself be regarded as a molecule of enormous size, and the result that the kinetic energy of any molecule will on the average be the same still holds for such a giant "molecule". The kinetic energy of the dust particle should therefore also on the average be $\frac{3}{2}kT$, and for the size previously given this works out to a speed of a little less than one centimetre per second.

The whole of kinetic theory was in fact developed long before there was any definite proof of the existence of atoms. While many of its conclusions were in agreement with the known facts, it contained no information about how large or small the molecules were, until it came to the study of Brownian motion. The comparison which I have just given allows a direct determination from the Brownian motion of Boltzman's constant and from this we can derive the number of atoms in a gas. Since the kinetic theory is much older than our knowledge of the structure of atoms this present chapter should have preceded the one on atoms, if we wanted to follow the historical sequence. I have chosen the opposite order because the evidence for the size of atoms which comes from Brownian motion and other applications of kinetic theory, involves a more complicated reasoning and is not as easy to visualize as the other methods.

MORE TESTS AND SOME DIFFICULTIES

We have so far only touched on the beginnings of kinetic theory, and even the more important of its applications are too numerous to be reported here. For example, it does give us a possibility of explaining the specific heat of gases. Consider a "monatomic gas" i.e. one in which each atom moves by itself without forming a molecule, as in the case of the "noble" gas helium. In heating it all we do is to make atoms move about, and therefore the kinetic energy $\frac{3}{2}kT$ is the whole of the heat energy which we have imparted to the gas. To find therefore the quantity of heat contained in such a gas we multiply $\frac{3}{2}kT$ by the number of atoms. The "specific heat", which is the increase of energy if the temperature is raised by one degree, is $\frac{3}{2}\,k$ per atom.

However, if we take a gas in which the molecules contain more than one atom, then, apart from travelling about, they are also capable of rotating about their centres. As a result of collisions they are in general in a state of rotation. Kinetic energy is also associated with this rotation. To find out just how much of it, we have to apply statistical considerations to the collisions of such molecules with each other and to the exchange of rotational energy with that of the motion of the whole molecule. The answer to this problem can be put by saying that the kinetic energy is the same for every "degree of freedom". We had introduced the idea of degrees of freedom at the end of Chapter 1 and found that the motion of a body, apart from rotation, involved three degrees of freedom, that of a thin rod five, if we did not include the rotation of the rod about its axis, or six if we did, etc. We must therefore take the kinetic energy per degree of freedom as $\frac{1}{2}kT$. A diatomic molecule like oxygen or nitrogen is like a rigid rod, but without the possibility of turning about the axis of the rod, since this would mean the atoms turning about their own centres. It therefore has five degrees of freedom and its energy should therefore be $\frac{5}{2}kT$. In other words, the specific heat of oxygen or air should be $\frac{5}{3}$ that of helium and this is found to be correct.

This argument depended on the oxygen molecule being perfectly rigid, i.e. on its two atoms being held at a fixed distance. In fact we would rather expect that this distance should be a compromise between attractive and repulsive forces. Two oxygen atoms will attract each other, according to the laws of chemistry, but if they get too close, i.e. come into actual mechanical contact with each other, there will be a repulsion, since the atoms are more or less impenetrable to each other. Hence there will be some distance at which the repulsion and attraction just balance each other and at this distance the atoms are in mechanical equilibrium, just like a weight suspended against gravity by a spring. If we pull it down too far the increasing tension in the spring will more than balance the downwards force of gravity on the weight; if it is too high the tension in the spring will be relaxed and there will be a net downward force. There is just one point where the forces balance.

Now if we do suspend a weight from a spring and hit it, it oscillates up and down until friction stops its motion. And in the same way we must expect that the two oxygen atoms in a molecule are disturbed from their equilibrium distance in the course of collisions and then carry out oscillations in which their distance from each other increases and decreases periodically.

If we apply our theory to the way in which collisions will set up such vibrations, we find again that there is on the average a kinetic energy of $\frac{1}{2}kT$ associated with this new degree of freedom, which corresponds to the vibration. But in this case the energy may be both in the form of kinetic energy and in the form of potential energy. To return to the example of our spring and weight, at the ends of the motion when the spring is extended or when the weight has been raised the greatest amount against gravity, it will just stop prior to reversing its direction of motion, and at that instant there is no kinetic energy; the whole of the energy is potential. It can be shown that on the average the two are equal and therefore the vibration also corresponds to an average potential energy of $\frac{1}{2}kT$. It would therefore follow that an oxygen molecule including its possible vibration contains not $\frac{5}{2}$, but $\frac{7}{2}$ times kT and this would not agree with the facts. Here we see a contradiction between the conclusions of kinetic theory and observation. This discrepancy is not always there. Some molecules do show this expected contribution from their internal vibrations. In particular in heavy molecules and at high temperatures, this is generally found. We shall see later that the discrepancy is not the fault of the kinetic theory, but is one of the indications that the basic laws of mechanics begin to fail on the atomic scale.

The same difficulty arises in a much aggravated form if we remember that each atom contains a nucleus and electrons. For example each oxygen atoms contains eight electrons, and if their motion is allowed for then the oxygen molecule should have not five, or six, but fifty-four degrees of freedom. If the collisions resulted in an irregular motion of the electrons inside the atoms we would be led to very much larger quantities of heat than are found in practice. The conclusion must be therefore that in the motion of a gas such as oxygen the internal motion of the atom is not affected by collisions. This is very different from what we would have to expect if some other star and its planets ever came close to our solar system; we find here another confirmation of the result already seen in the last chapter, that there are important differences in the behaviour of an atom from that of the sun and planets.

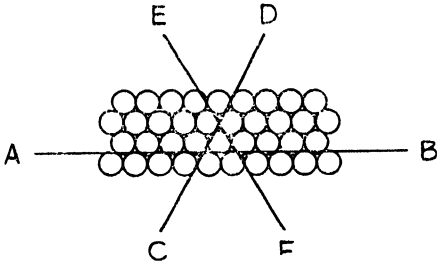


Fig. 27. Close-packed spheres

While we are therefore for the present unable to attempt a mechanical description of what goes on inside the atom, these results force us to regard, for most purposes, the atom as a rigid structure, whose internal motion is not easily affected by forces from outside.

SOLIDS AND LIQUIDS

So far we have applied the kinetic theory to the behaviour of gases, but it can also be used to understand the structure of solids and liquids. To start with solids: we had already seen that chemistry required attractive forces between atoms in some cases, but also that this attraction would change into repulsion when the atoms actually came into contact with each other. If then I have a large number of atoms in a small space the forces try to get the atoms as close together as possible without their actually overlapping, and this leads in general to some rather regular pattern. For example, if we imagine that the atoms behave as hard spheres of equal size which all attract each other, then they will tend to arrange themselves in "close packing", as in Figure 27. If we fill a shallow box or tray with small equal spheres, a child's marbles for instance, or steel balls from a ball bearing, and shake the box a little to allow the balls to settle down to their most compact pattern they will arrange themselves in layers, each of which looks exactly like Figure 27.

Actual atoms may form different patterns because the forces between them are different. They may be softer (resembling rubber balls rather than marbles) and attractive forces may play an important part. In a compound, different atoms may have different size, and this, too, affects the pattern. But the interesting point is that it always comes out to be some regular pattern in which some basic group of atoms is repeated periodically. Such a regular pattern of atoms is called a "lattice", and most solid substances are made of such lattices of atoms. This regular repetition makes them perfect diffraction gratings for the diffraction of X-rays which we discussed in the previous chapter, and in fact one can use the diffraction of X-rays to determine the pattern of atoms in the crystal lattice.

A regular pattern like that of Figure 27 favours certain directions over all others. For example if a structure like that is broken into two this is easiest along a line such as *AB* or *CD* or *EF*. Therefore such "crystals" will tend to break in such a way that their faces form certain definite angles with each other. This behaviour is familiar in the case of certain minerals and stones and in particular for all precious stones; grains of salt or sugar also show this tendency clearly. In metals such as steel or aluminium it is not evident, because these normally consist of a mosaic of very small pieces, each having a pattern like Figure 27, but with the patterns in different pieces turned in different directions. Such a structure is called microcrystalline, made of small crystals, and in metals the nature of these micro-crystals and their size is of great technological importance.

That atoms with attractive and repulsive forces between them will tend to take up such regular patterns, is an immediate consequence of mechanics, though we cannot foresee what will be the pattern in any given case without knowing more about the forces than we have learnt so far. What happens now if the substance is hot? In that case again the atoms are in motion, but in the solid they cannot run away from each other because of the forces between them. The result is that the atoms vibrate about their normal positions. To find out how strong this vibration would be we have again to apply kinetic theory, and it turns out that the rule about the average kinetic energy is still correct. If there are N atoms in our solid substance, then each has three degrees of freedom and therefore their kinetic energy should on the average come to $\frac{3}{2}NkT$, and again the potential energy on the average equals the kinetic so that the total energy of the vibration comes to $3NkT$. The way in which this is proved in kinetic theory or statistical mechanics is somewhat abstract, but it may also be confirmed by considering the collisions of gas molecules with the surface of such a crystal. It then turns out that on the average the gas molecules neither lose nor gain energy in such collisions if their kinetic energy on the average equals the kinetic energy of each of the atoms in the crystal.

As a result we conclude that the quantity of heat contained in a solid substance at temperature T is $3NkT$, where N is the number of atoms. The specific heat is $3Nk$. This "law of Dulong and Petit" agrees well with observation on most solids, provided the temperature is not too low, but in all of them it fails when we cool them far enough. The amount of heat contained in a solid near absolute zero is very much smaller than corresponds with the law. Its behaviour at low temperatures is thus another contradiction which gives warning that our statement of the laws of nature is as yet incomplete.

In a similar way one can see that the elasticity of solids, i.e. the fact that force is required to alter their shape and that they will tend to return to their normal shape, can be understood in terms of atoms and the forces between them.

Their strength, i.e. the force required to break them or deform them permanently represents a much more difficult problem and it is only in the last twenty years or so that it has become possible to understand this in quantitative terms.

We can understand that solid substances will melt, i.e. become liquid when heated to a sufficient temperature. As the temperature rises the vibrations of each atom about its normal position become more violent. Also the neighbours of any given atom will deviate more and more from the regular positions where they belong in the pattern. Now the regular position of each atom is determined only by the forces from its neighbours and if its neighbours are disturbed from their regular positions then the tendency of the atom itself to be in the right place is also weakened. Therefore as the amount of vibration of the atoms about their normal position increases, their tendency to form a regular pattern decreases, until a point is reached where the regular order collapses altogether and the arrangement of atoms becomes irregular.

We may perhaps compare this with a long column of soldiers marching in step. As they get tired a few of them do not keep in step properly. This will not upset the regularity of the marching column until the point has been reached where, within the range of vision of each soldier there are as many others out of step as there are correctly in step and at that point all order will have gone.

This essentially is the process of melting. We may also understand from the atomic picture how a liquid or a solid can turn into a gas. Consider the atoms near the surface of a solid, or a liquid. These atoms are vibrating, but they always remain under the attractive force of their (regularly or irregularly spaced) neighbours, but the velocity with which the atom moves about is not always the same; sometimes it will happen to be slow, sometimes it will vibrate more strongly because it so happens that the movement of all its neighbours combines to push it in a certain direction. From time to time it may happen that an atom gets up enough speed to get away from its neighbours out into empty space. In this way the space above the liquid or solid will contain some atoms which are moving about freely as a gas. Occasionally, such a gas atom strikes the liquid or solid surface and shares out its kinetic energy with the atoms already there. As a result it usually does not retain enough energy to get away again. If we start with only the solid or liquid and no gas at all then atoms go on escaping from the surface to increase the amount of gas, until the number of atoms hitting the surface and getting caught equals the number leaving. Now the number of atoms which get away from the surface evidently depends on the temperature. The faster the vibration of the atoms the greater will be the chance that an atom can pick up enough energy to get away, and we therefore see that the steady vapour pressure above the solid or liquid, i.e. that pressure for which as many atoms evaporate as condense again, increases with increasing temperature.

These examples will suffice to show that all the behaviour of matter, as we normally know it can, in the last resort be reduced to the mechanical behaviour of atoms and to the laws of probability or statistics applied to large numbers of atoms. Needless to say, we have not in this chapter reached a stage where we could in fact predict the actual mechanical characteristics of any given substance from first principles, since we are not in a position to know precisely what the forces between atoms are. But the important point is that we see these problems reduced to the problem of atomic structure and the interaction forces between atoms, so that we can recognize that the behaviour of matter is a consequence of fundamental laws which govern the structure and the behaviour of atoms.

While in very many ways our reasoning based on Newton's mechanics and the kinetic theory has been confirmed, at the same time there remain important contradictions, and we cannot claim to have understood the structure of matter until these are cleared up.

RADIANT HEAT

A particularly important contradiction of this kind arises when we consider the emission of electromagnetic waves, such as heat radiation or light, by a hot body. We are very familiar with this process in practice, and it can be watched if for example we switch on an electric fire and allow the element to heat up gradually. At first it produces no visible light, but by placing our hand near it we can feel that it sends out radiant heat, which, as we have already seen, consists of so-called infra-red rays. As it gets hotter the intensity of this radiation increases and eventually it shows a faint red glow. We then say that the element is "red hot". As the temperature rises further the glow becomes more intense and at the same time changes in colour more towards the yellow. For the normal electric fire that is the highest temperature we can reach, but we may go on to look at the filament in an electric light bulb, which is just the same kind of wire as in the electric fire, only hotter still. Here the light is much more intense and has almost the same colour as sunlight. We say the wire is "white hot". In the still hotter arc of an arc light, the light has a bluish tinge, i.e. the colour has shifted even further towards the blue and violet end of the spectrum than in sunlight. A carbon arc may be hotter than the surface of the sun. It is actually found that the colour and intensity of the light does not depend much on the nature of the hot body by which it is produced, but mainly on its temperature. The ideal case is that of the so-called black body, and it is one of the key problems in the science of heat to explain the amount of radiation coming from a black body of a certain temperature and its colour.

This question may also be stated in the following way. Imagine such a black body contained in a space whose walls are perfect mirrors and therefore reflect the radiation without absorbing any of it. Then as more and more radiation is emitted by the hot body, more and more of it will be travelling back and forth between the various reflecting walls, except that some of it will fall on the body again and be absorbed. This will go on until a steady state is reached, in which the body absorbs as much radiation as it produces. If we know how much light of a given colour there is in this steady state, we can work out from it also how much the body emits in any given time.

Now the application of kinetic theory to this problem gives a perfectly simple, but in its consequences surprising answer. Namely the answer is again that the electric and the magnetic energy of the light waves should each amount on the average to $\frac{1}{2}kT$ for every degree of freedom. But how many degrees of freedom does the radiation field or the electromagnetic field in such an enclosure comprise? Since to specify the electric field completely we must know its strength and direction at *any* point in space, and since the number of points in our enclosure about which we may ask this question is not limited, it would follow that there exist an infinite number of degrees of freedom and therefore no steady state should ever be reached. This is contrary to observation.

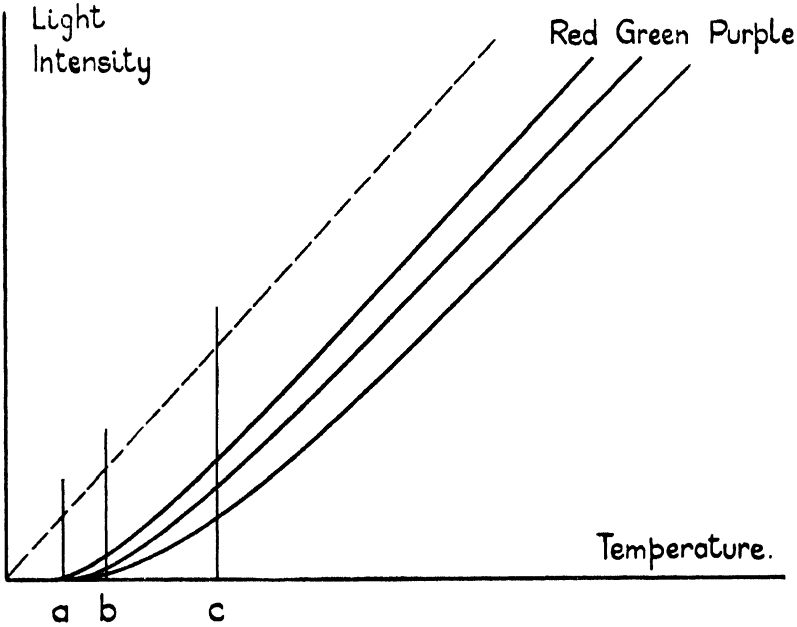


Fig. 28. Heat radiation.

The situation is further illustrated by Figure 28. This is a plot of the intensity at different temperatures of the light of some particular colour. The temperature is plotted horizontally and the light intensity vertically. The broken straight line is the result of the simple theorem from kinetic theory which predicts that the radiant intensity of any colour should rise proportionally with the temperature. The various full curves show how the light intensity actually does vary for light of different colours. At very high temperatures each of those curves runs parallel to that of the simple kinetic theory and falls short of it only by a relatively small amount. But at low temperatures the intensity actually found is very different from that predicted. For instance at the temperature corresponding to the vertical line marked *a* there is practically no visible light at all, although kinetic theory would predict an appreciable amount. At *b* there is a fair amount of red light, but much less green and no violet. At *c* all colours have comparable intensity.

Here we have again a case where the apparently certain prediction of the kinetic theory fails. We might simply add that to the list of such discrepancies already noted, but this particular one is of special historical importance, because in thinking about this discrepancy Planck was led to introduce the hypothesis of the quantum of action, which opened up the way to the solution of most of these difficulties.

These ideas will occupy us for a good part of the remainder of this book, but in the course of their discussion it will be convenient to refer to the results of another important development of the twentieth century, the theory of relativity.

6

Relativity

GENERAL REMARKS

In this chapter we shall meet the first of the revolutions in scientific thinking which have been brought about by the discoveries of the 20th century. In the jargon of the physicist the contents of all the preceding chapters are often referred to as "classical" physics, to distinguish them from the "modern" theories which follow.

The ideas of relativity at first met with rather strong opposition both amongst physicists and amongst philosophers. The physicists met the new hypothesis in the critical spirit in which it is their business to regard any new idea, until it has passed the triple test which physics requires. It must firstly leave undisturbed the successes of earlier work and not upset the explanations of observations that had been used in support of earlier ideas. Secondly it must explain in a reasonable manner the new evidence which brought the previous ideas into doubt and which suggested the new hypothesis. And thirdly it must predict new phenomena or new relationships between different phenomena, which were not known or not clearly understood at the time when it was invented. This process took some time because relativity is important only for objects moving with a speed comparable to that of light. As such objects were not readily available the opportunities for tests were few, and in many cases the tests required very difficult observations of high precision. Since then particles moving with high velocity have become commonplace in any physics laboratory. We are no longer concerned with small corrections in the behaviour of these particles which require measurements of high precision, but the relativistic features of their motion have large effects which it is quite impossible to overlook. To put the point in its crudest form: in developing machines for physical research in which particles are accelerated to very high speeds, engineers have to incorporate devices, costing many thousands of pounds, which are required only because of the relativistic features of the particle motion. They are scarcely inclined to regard such features as a result of idle or mistaken speculation. To-day no physicist who has practical knowledge of work with fast particles would question the principles of relativity.

The opposition from philosophers arose because the theory of relativity called in question statements that had been regarded as the concern of philosophers. It was doubted whether the physicist had the right to query ideas which the philosophers had regarded as evident and unquestionable truth.

By now it appears to be widely recognized that our ideas about space and time are derived ultimately from our experience of the outside world and that many statements which we regard as evident are true only within the limitations of our practical experience. They may turn out to be unjustified prejudices when extended to situations not normally familiar to us.

In the history of physics such situations are common. The discovery that the earth was revolving about its own axis and about the sun ran counter to mechanical intuition built on simple experience. When we learn at school about the opposition first encountered by this idea we tend to regard the scholars of the Middle Ages, who failed to accept it, as incredibly narrow-minded in their outlook, but the only reason why we ourselves find the thought easier is that we have become familiar with it through our early education and have accepted it before our critical power was developed far enough to question it seriously.

The idea of Galileo's law that a moving body tends to continue in its motion runs counter to everyday experience and was at first quite legitimately questioned by the scientists of the time. It was accepted only with difficulty by thinkers who generalized abstract ideas about motion.

The fact that light travels with finite speed is another fact which runs counter to our intuition. It is hard even to get used to the fact that the speed of sound is finite, and even the simple experience of watching a man chop wood at a distance, when the sound of the blow reaches us later than the sight of the falling axe, requires a conscious adjustment of our senses to an unfamiliar situation. We have learned even more strongly to rely on our eyes as telling us what goes on in the outer world and to regard the information they give us as true information about what goes on at the instant we see it. There is nothing in everyday experience on the surface of the earth which will disprove this illusion, but three hundred years of experience in astronomy and more recent methods for measuring times to enormous precision have got us to accept the finite speed of light.

The wave nature of light causes similar difficulties for our intuition.

Only a conscious effort of reasoning makes us accept the fact that diffraction of light waves at a sharp edge can make the light actually travel round a corner.

Thus every new development of science requires us to drop some deep-seated prejudice, but probably all the earlier developments, with the exception of Galileo's law, were made easier by the fact that we could explain them in terms of familiar mechanical models. For example, the experience of the finite speed of sound prepared us to accept the finite speed of light. The behaviour of light waves can be visualized by observations of the ripples on a pond, the difference being only one of scale, but such aids fail us in getting used to the ideas of relativity.

The acceptance of relativity was probably delayed by its name, which suggested a superficial connection with the philosophical concept of relativity, according to which all truth was regarded as relative. As we shall see, nothing is further from the contents of the new development. In relativity, the laws of physics have a precise and absolute form, only certain specific statements that our intuition leads us to regard as absolute, turn out to be prejudiced.

MOTION AND REST

Long before the developments which led to Einstein's formulation of the theory of relativity it was known that many of the laws of nature, in particular the laws of mechanics, did not alter their appearance if the observer, instead of standing still, was moving with uniform velocity and direction.

It is well known that the passenger in a railway train is not aware of the motion of the train unless the vibration of the coach, due to the roughness of the track, or the acceleration caused by starting or slowing down, or by the curvature of the track, show him that the train is moving. Frequently, when we look out of the train window at another train in the station we are confused as to whether it is our train or the adjacent one which has started moving. The question is settled when we catch a glimpse of the station buildings or other objects, which, we know from experience, usually stand still. If a fellow passenger asserted that in fact the station, with the track and all the landscape, was moving and the train was standing still, and if he was not impressed by the argument that stations and landscapes don't usually behave like that, we would not find it easy to prove him wrong. Any mechanical experiment which we carry out inside the compartment would proceed precisely in the same way as if the train was standing still (apart from the vibrations and curves to which reference has already been made). And this fact is usually expressed by saying that the laws of mechanics are the same for two observers who move relatively to each other with a uniform velocity.

This is just as well, because it would otherwise have been very much harder to develop the laws of mechanics working on the surface of the earth, because, as we now know, the earth is in fact not standing still, but, with us and our laboratory and all our measuring apparatus, is turning about its axis, moving us through space at a speed of roughly a thousand miles per hour. This is not uniform motion, but motion in a circle, and therefore, as we had seen in Chapter 4, gives us also an acceleration. But, as the period of revolution is 24 hours the change in direction is slow. While the speed is very high, compared to speeds to which we are accustomed in practice, the acceleration is very small. The acceleration can be shown by observations, for example, on the so-called "Foucault pendulum", but the speed itself has no effect on any mechanical process.

Apart from this, the whole earth is moving on its orbit around the sun with even greater speed, and the sun with all its planets is itself moving in relation to other stars.

While the laws of mechanics therefore do not allow us to distinguish uniform motion from rest, this would appear to be different when our description of nature includes electricity and light. In discussing the laws of the electromagnetic field in Chapter 2 we found that electromagnetic waves, which include light waves, always travel with a fixed velocity, which we called c. We should therefore expect to find some difference in the apparent speed of propagation of light, owing to the fact that we are moving ourselves. For instance, if the only motion was due to the earth's rotation, light from a lamp placed to the east of us should reach our eye more quickly because our eye is in fact moving towards the lamp and therefore coming to meet the light waves. Similarly light from a lamp placed in a westerly direction should appear to travel more slowly because it has to overtake our eye, which is moving away. If u is the velocity with which the surface of the earth is moving, the apparent velocity from a light source in the east should in fact be $c + u$ and that from a light source in the west $c - u$. The difference is of course very small, because u is about a million times smaller than c, but careful experiment ought to reveal the difference.

This expected result was sometimes called the ether wind, because one thought at that time that all space was filled with a hypothetical substance called ether which acted as a carrier for electromagnetic waves. The effect of our motion through the ether on the propagation of light was similar to the effect of a strong wind on the propagation of sound, when it is well known that sound will appear to travel more rapidly down-wind than up-wind.

In their famous experiment Michelson and Morley set out to detect this "ether wind". To avoid the difficulty of measuring distances and times to the required high precision they used in fact a light beam which was split in two. One part was travelling, say, in an east-west direction and back after reflection from a mirror, and another north-south to another mirror and back. On being combined again these two light rays would form interference fringes of the kind we have discussed in Chapter 3. The position of these fringes depends on how many oscillations each light wave has undergone during its travel. It is not difficult to work out the times taken for the return journey by either light beam. This calculation, which we shall omit here, shows that the effect of the earth's motion is the same as if the east-west arm of the apparatus was lengthened in the ratio of 1 to $\sqrt{1 - \frac{u^2}{c^2}}$. It is difficult to measure the length of the light paths to the required accuracy, but Michelson's reasoning was that on turning his apparatus through a right angle, so that now the other arm pointed in the east-west direction, this was bound to lengthen the time of travel of the one light ray and shorten the other, with a resultant shift of the interference fringes. Since the earth not merely rotates, but also revolves in its orbit about the sun, the effective speed with which we are moving should also alter its direction between day and night. Therefore even if the apparatus was left still, the fringes should alter their position in the course of a day.

The result of this experiment was completely negative. Whichever way the apparatus was turned and however long one looked at it no shift in the fringes was observed. All other attempts to observe the "ether wind" failed similarly. And hence the one kind of experiment which appeared to make it possible to distinguish motion from rest had failed.

THE LORENTZ CONTRACTION

It took a long time to understand completely the implications of this result. Various explanations were tried. It was for example suggested that the velocity of light might be dependent on the motion of the light source. Since the Michelson experiment employed a lamp which was fixed on the earth's surface one would then expect the light to travel with a given speed relatively to this source, i.e. relatively to the surface of the earth. But this explanation had to be given up on reflection, since we know that the light from the sun, for example, really comes from atoms in the atmosphere of the sun, which according to the results of the preceding chapter are moving about irregularly and with great speed. The light from different atoms should therefore, if this explanation were right, have different speeds. If we looked not at the sun but at a distant star, which is similar to the sun, but further away, we should see this star not as a point, but as a streak, since light from different atoms on the star should have come to us with different speeds and should therefore have started when the star was in quite different positions in relation to our telescope.

It was also suggested that in the neighbourhood of the earth the ether should take some part in the motion of the earth, so that there was an "atmosphere" of ether around us, which moved in the same way as the earth itself. This hypothesis, too, was untenable, because if it were true there had to be somewhere outside the earth a transition region where the ether's speed changed from that of the earth to that of free space. This would lead to a refraction of light rays and an apparent displacement in the position of stars, which is not observed.

More such alternatives were considered and found equally unsuccessful. Nearer to the truth came a suggestion by H. A. Lorentz, which at first sight seems very far fetched. Lorentz suggested that every moving object contracts in the direction of motion in the ratio of $\sqrt{1-\frac{u^2}{c^2}}$ to 1, if u is its speed. If the arm of Michelson's apparatus which points in the direction of motion had in fact contracted by that amount this would just make up for the effect of the "ether wind". When one first hears of this "Lorentz contraction" it sounds most artificial and unreasonable. It sounds, in fact, as if nature had conspired to adjust its laws in such a way as to make it hard for us to observe the ether wind. There seemed no reason in mechanics why bodies should undergo this peculiar contraction.

However, there was a much better basis for the suggestion than appeared at first sight, because Lorentz knew already that all matter consists of atoms, as we had seen in Chapter 4, and that atoms contain positively and negatively charged particles, the forces holding them together being electric forces. Now in the laws of electricity and magnetism the velocity of light occurs as a constant. If charged particles move with a speed not very small compared with that of light, then their electric fields become more complicated. A point charge, for example, such as that of an atomic nucleus, has an electric field in which the potential depends only on the distance from the charge, so that all points on a sphere centred on the charge are at the same potential. If the charge is moving it produces not merely an electric, but also a magnetic field, because the transport of electric charges means the same thing as an electric current. In addition the electric and magnetic fields at any point in space near which the charge passes vary with time. We saw in Chapter 2 that complications arise in the laws of the electric and magnetic field if the fields vary in time. This more complicated mathematical problem can be solved, and the result is that the surfaces of constant potential are no longer spheres, but are flattened in the direction of motion, by just the same amount that would be required by the Lorentz contraction.

This change in the electric forces no doubt would cause a change in the size and shape of an atom and in the distances of different atoms from each other. Not enough was known at the time about atomic structure to be sure what the overall effect would be. At least the possibility emerged that this Lorentz contraction could just come out as a consequence of the effect of the motion on the electric forces which hold atoms together.

This idea of the Lorentz contraction would, if confirmed, account for the negative result of the Michelson experiment, but at first sight it would open up a new possibility of finding out whether a body was at rest or in motion. Consider again for simplicity only the motion due to the rotation of the earth, and take a very accurately measured cube, which has one of its edges in the north-south direction. If we turn it through a right angle, the edge which was in a north-south direction before, and now is in an east-west direction, will have undergone the Lorentz contraction, whereas the other horizontal edge which previously was east-west and therefore was contracted, will have expanded. In the new position measurement should show us therefore that the body was no longer a cube. Now first of all, the amount of the Lorentz contraction is so small that it would be extremely hard to detect the change, but we can imagine that we had measuring instruments of sufficient accuracy to do so.

We should, however, then find a much more fundamental difficulty. The only way to measure a length is by comparing it directly or indirectly with a measuring rod or tape. This measuring rod itself, like any other rigid body, would be subject to the Lorentz contraction and therefore, as we turn it to place it first against the one edge and then against the other edge it would follow the Lorentz contraction.

The two edges would still appear to be of the same length in terms of the marks on the measuring rod.

If we therefore apply consistently the postulate of Lorentz to all solid bodies without exception then it would be impossible in principle to detect the contraction by any direct measurement.

HOW TO COMPARE LENGTHS AND TIMES

But one might think that the Lorentz contraction would still be a suitable way to find out which of the two passing trains with which we started this discussion, is standing still, and which is moving. Suppose that the coaches of both trains are of identical construction, then that of the moving train should have contracted and therefore be shorter than that of the stationary train. To make matters easier imagine for the moment that the speed of the moving train was not one-millionth, but one-third of the velocity of light; then the Lorentz contraction would shorten it by five per cent, which should be easy to see. Thus all we would have to do is to watch carefully as two coaches pass each other and to see which is the longer.

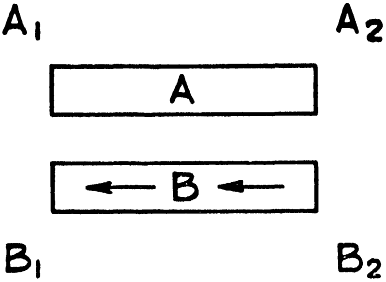


Fig. 29. Passing trains.

However, this also is much more involved than it seems, if we remember that we are dealing with very high velocities and very short times. To make this point clear consider Figure 29. *A* is meant to indicate the stationary coach and *B* the moving one. If *B* is really shorter, then at the instant when its front end B_1 is in line with the end A_1 of the stationary coach, its rear end B_2 will already have passed the end A_2 of the stationary coach. But to be sure of this we must be very sure of our time. If we look at B_2 a very little too early it will not yet have passed A_2 and we would then think the coach too long. In other words, comparing the lengths of the two coaches boils down to being able to tell with great accuracy the time of passage of the ends past each other. The moving coach is longer, or shorter, than the other if its end B_2 passes A_2 after, or before, the end B_1 passes A_1.

We may, of course, look at the two ends simultaneously, but we then get involved in complications due to the finite speed of propagation of light. Supposing we stand at A_1. At the time when we see B_1 passing, we shall also see the end B_2, not as it is at that instant, but as it was a very short time earlier, namely earlier by the amount of time which the light has taken to travel from B_2 to A_1. It is true that we can correct for this, since we know the speed of propagation of light, but in working out this correction we must firstly know the length of our own coach, i.e. we must assume that our own coach is standing still, and thus no Lorentz contraction has to be applied, and also that the light travels with its normal speed and no ether wind has to be allowed for.

We might try to get some time signal across from A_2 to A_1 in some other way. We could imagine that an assistant stationed at A_2 presses a button when B_2 passes and that this gives some kind of signal at A_1, but this transmission must either be by mechanical means or by an electric current or some other device, and all these means of communication are in fact slower than the speed of light. To work out their speed of propagation with the required accuracy gets us into the same difficulty.

Let us then limit ourselves to light signals and consider how the whole process would appear from the point of view of an observer on the other train who is trying to apply the same test. He would also watch the passage of A_2 past B_2 and A_1 past B_1 and he would also transmit a light signal from one end of the coach to the other and apply a correction for the time this has taken. He would naturally assume that his train was at rest and would apply no correction for ether wind or for the Lorentz contraction. The interesting result of working this out in detail is that on the same basis he would come to the conclusion that our coach has been shortened by the Lorentz contraction, i.e. that A_1 and B_1 are in line *before* A_2 is in line with B_2.

The important point to realize is that the observations of both observers are completely interchangeable. We are coming to the conclusion that the coach B is moving and therefore shortened by Lorentz contraction, whereas the other observer, starting from the assumption that his coach is at rest, would conclude that our coach A is moving and therefore shortened. No doubt we will criticize his measurement because he has made assumptions about the length of his coach and about the absence of an ether drift which to us appear wrong, but he will answer by pointing out that from his point of view the same criticism applies to our method.

The essence of the argument, which is a simplified version of that used by Einstein, is that the length of an object has no absolute meaning, but depends on the state of motion of the observer by whom it is determined. Similarly the question of time, i.e. which of two events occuring at two distant points is the earlier, has no absolute meaning, but depends on the point of view of the observer.

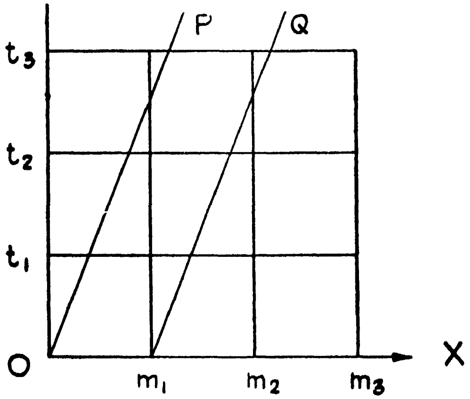


Fig. 30. Distance-time graph.

These ideas certainly are in conflict with our intuition, which regards the length of an object or the time at which an event takes place as absolute concepts, not necessarily related to any actual process of measurement. One has to get accustomed to the realization that the certainty with which we take these things for granted comes only from the experience of everyday life. In practice any doubt about the meaning of the exact length of a moving object, such as a railway coach, is completely negligible within the limits of error of any practical determination. In practice also we never get into trouble by assuming that any event which we can see takes place when we see it, since for practical purposes the speed of light may be regarded as infinitely large.

It is one of the merits of Einstein's theory to have brought out clearly the fact that our ideas of space and time are abstractions from experience, and that in applying them to situations involving greater speeds or shorter times than those to which we are normally accustomed we cannot take their meaning for granted without specifying them precisely. This is done only by formulating questions which can, in principle, be answered by an actual observation.

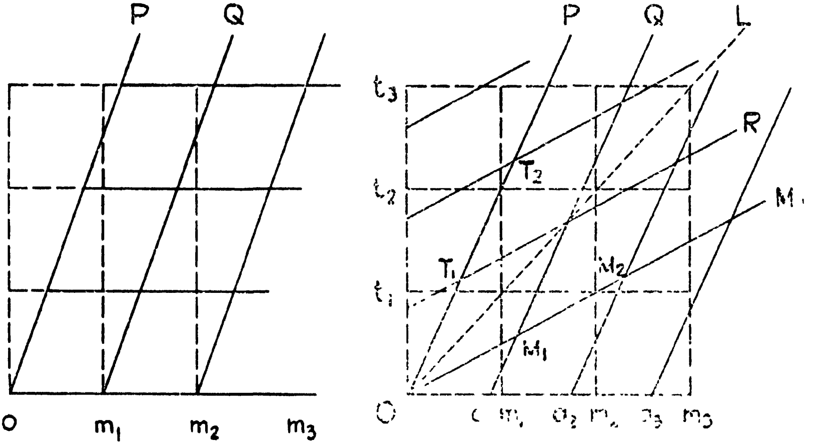
So far we find, then, that the postulate of the Lorentz contraction is consistent with the principle that the laws of physics are precisely the same for two observers moving uniformly with respect to each other and that there is no observation by which the point of view of one observer can be preferred over that of another.

The connection between the observations of different observers may be expressed by means of diagrams. Figure 30 shows a kind of graph in which distances are plotted horizontally and time vertically. The horizontal line OX shows some line, for example the railway track under discussion, at some particular instant which we take as the zero of time. Along this we have marked out equal distances with "milestones" m_1, m_2, m_3, etc. The vertical line Ot indicates the passage of time at the starting-point O and equal time intervals have been marked off at t_1, t_2, t_3. Subsequent horizontal lines show the track again at the later times and subsequent vertical lines the passage of time at the various milestones. Such diagrams are in fact used in constructing timetables, since one can indicate the passage of a train in such a diagram by a sloping line, such as OP or m_1Q and one can then tell at a glance at what time the train will pass a particular point, or where two trains will meet. In our case OP and m_1Q are meant to show the two ends of one coach of the moving train.

If the observer on the moving train also sets up distance marks on the train and records times, how would his finding compare with ours? For the case of a real train, when all velocities are small compared to that of light, and relativity is therefore unimportant, this is represented by Figure 31a. The sloping lines now indicate points which to the observer on the train appear to stand still and which he regards as marking out distances; indeed the end of each coach is to him a fixed landmark. Our own milestones are shown again as broken lines for comparison; the time marks are the same for both of us, since apart from relativity the time is the same for all observers.

Compare with this the relativistic picture which arises when we are dealing with really high velocities. For this we must go to a different scale and in Figure 31b the situation is drawn again, assuming that our "milestones" are so far apart that light will just go from one to the next during one time interval. (For instance, if the time interval is one second, the distance between the milestones must be 300,000 kilometres.) On this scale the motion of any real train would appear as a practically vertical line, but we have shown in the drawing the case of a "train" moving with one-half the speed of light.

Again OP and a_1Q show the movement of rear and front of one coach. a_1 now does not coincide with m_1 because, owing to the Lorentz contraction, the coach appears to have shortened, and at one instant of time it occupies less than the distance between O and m_1. However, the time marks of the moving observer are now given by the sloping lines like $O\ M_1\ M_2\ M_3$, etc., and therefore to find the length of our coach the moving observer would see how much of this line is cut by the lines indicating the ends of our coach, i.e. the two vertical broken lines through O and m_1. The second vertical line passes to the left of M_1 and the moving observer will therefore find our coach shorter than this.



(a) (b)

Fig. 31 (a). Moving observer; low speed.
(b). Moving observer; one-half light velocity.

So far the diagram only repeats in graphical form what we have discussed in words already, but it also permits a further conclusion. Consider the dotted line OL. This indicates the passage of a light signal, since a movement along this line describes a motion which takes just one of our time intervals to travel to the next milestone, and this by convention just represents the velocity of light. Now we know from the Michelson experiment that the velocity of light must appear to be the same for the moving observer and therefore the line OL must also pass through the corners of his network. The diagram has been drawn in accordance with this, and we see that the later time marks of the moving observer such as T_1R will intersect our time mark between Ot_1 and OP.

From this we can deduce the effect of motion on the measurement of time, i.e. on a clock. If the observer carries a clock which passed our stationary clock at O and was then synchronized with it, it will "strike", i.e. indicate the passage of exactly one time interval at T_1, but to us this is later than t_1. We therefore conclude that the clock carried by the moving observer is slow. From the symmetry of the figure it is slow compared to ours in the same ratio in which the distances have been contracted by the Lorentz contraction. This effect on moving clocks is known as the time dilation. It is again true that with exactly equal right the observer on the train will conclude that our clock is slow, since its "stroke", t_1, lies above the line T_1R, i.e. occurs from his point of view after the end of the first time interval.

This predicted time dilation can be linked very directly with the known behaviour of fast particles, as we shall see in Chapter 11.

COMPOSITION OF VELOCITIES

With this new relationship between the records of time and distance kept by different observers there must also go a new rule for the composition of velocities. If a man on a train moving with velocity u fires a gun in the forward direction, which gives the bullet a speed v with respect to the train, then the total velocity of the bullet as seen from the track would be $u + v$. But in relativity we must translate the path of the bullet, which to the observer on the train is just motion with velocity v, by correcting for the difference in viewpoint about times and distances and if this is carried out we find the result $\dfrac{u + v}{1 + \dfrac{uv}{c^2}}$. If u and v are both small compared to c, the velocity of light, then the second part of the denominator is very small compared to 1, and the result is again that the velocities just add. But if u and v are not so small then the answer is different. For example if u and v are each one-half of the velocity of light the combined effect of the two velocities is $\frac{4}{5}\,c$. One sees easily that if u and v are both less than the velocity of light, however close they may be to it, the resultant velocity will still be less than c.

This rule is correct only when the two velocities are in the same direction. For the more general case the rule is more complicated and I shall not give it here.

One particular consequence of this law of composition of velocities is evident. If $v = c$, i.e. if instead of firing a gun the man on the train sends out a light signal, then the combined velocity is again equal to c, and this brings us back to the starting point, namely that light appears to be travelling to both observers with the same velocity.

This principle of the constant velocity of light is the most basic law of relativity. To avoid misunderstanding it should be stressed that in all this when we have spoken of the velocity of light we ought really to have said "the velocity of light in empty space". In any material medium, even in air, and to a greater extent in such dense substances as water or glass, we know from the refraction of light, which we discussed in Chapter 3, that the velocity of light is different. This is caused by the effect of the light wave on the electric charges contained in the atoms. A light ray passing through glass will therefore not appear to different observers as propagating with the same speed. This is not surprising since the situations seen by the two observers are then different. For the one, glass or water is at rest, whereas for the other one it is moving. It was in fact in a study of the propagation of electromagnetic waves through moving bodies that Einstein was first led to a clear understanding of the arguments I have sketched above.

MECHANICS OF FAST-MOVING OBJECTS

This new application of the properties of distances and times also requires us to change the laws of mechanics as applied to fast-moving bodies; to see what modifications are required one can reason in the following way: look at the second law of Newton in the form that the force equals the rate of change of momentum. Before we can generalize this law we must be clear that both the ideas of momentum and force may need revision; the form which the mechanical law will take depends to some extent on which of the two, if either, we want to keep unchanged.

It turns out that the clearest statement of the laws is obtained if we retain the law of action and reaction, which says that the forces which two objects exert on each other must be opposite and equal. In consequence, the interaction between two objects does not change the total momentum. We also want to retain the law of conservation of energy, and in particular of mechanical energy, if we are dealing with an elastic collision. Consider then the collision between two elastic identical objects, for example two billard balls. If these collide with equal and opposite speeds their combined momentum must be zero, and if momentum is to be conserved it must be zero even after the collision. This means after the collision the two billard balls must still travel with opposite and equal velocity.

In Figure 32 we have shown two such balls colliding, with the arrows indicating their intitial velocity of approach; the equal length of the arrows is meant to indicate that their speeds are equal. Now suppose that they have collided in such a way that they separate again in a direction at right angles to the original. This is indicated in the figure by the broken arrows. The speeds must again be equal.

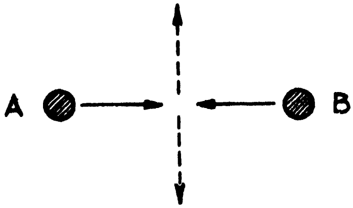


Fig. 32. Collision of two objects.

Moreover, energy conservation requires the speed with which the balls separate to be equal to that with which they collided: If the velocity of approach is u the total kinetic energy before the collision is twice the kinetic energy of a ball moving with velocity u, and if they separate again with velocity u this will also be the final energy. Note that this argument is valid regardless of how the kinetic energy depends on velocity.

We conclude that the situation sketched in Figure 32 is mechanically possible, i.e. preserves momentum and energy. If the mechanical laws are to be the same for all observers, the laws of conservation of energy and momentum must also appear satisfied if this same process is seen by a moving observer. Supposing that the moving observer is travelling with speed v from right to left, so that to him the ball A will appear to travel faster and B more slowly than before; after the collision both will apparently not be travelling at right angles to the original line, but more towards the right. This is shown in Figure 33. The important thing is that from the law of composition of velocity we can work out the directions and magnitudes of all the velocities as seen by this observer, i.e. the directions and lengths of all the arrows of Figure 33. Now the momentum and energy must still be conserved, and this tells us something about the

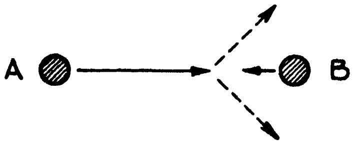


Fig. 33. The collision of Fig. 32, seen by a moving observer.

momentum and energy of objects moving with various velocities. I shall not give the details of this calculation here, which requires a little more mathematics than we are using, but from it the dependence of energy and momentum on the speed of a particle can be determined quite unambiguously.

The result is that the momentum p of an object of mass m moving with velocity u is $p = \frac{mu}{\sqrt{1 - \frac{u^2}{c^2}}}$. This relation between p and u is shown in graphical form in Figure 34. The full line shows the relativistic relation, whereas the broken line represents the law of Newtonian mechanics, where $p = mu$. It is seen that the relativistic momentum for small u is practically the same as before, but for velocities near that of light, very much larger than the momentum of Newtonian mechanics. This same result can be expressed differently; we may still retain the definition of momentum as mass times velocity, provided we are prepared to regard the mass of the object as variable. If we denote by M the quantity $M = \frac{m}{\sqrt{1 - \frac{u^2}{c^2}}}$ then indeed the momentum is $p = Mu$. To distinguish the two definitions of mass one sometimes calls m the "rest mass" of the object, i.e. the mass which we determine by dynamical measurements when the object is at rest or moving very slowly. Even if we work with this idea of a variable mass we must, however, remember that we should not

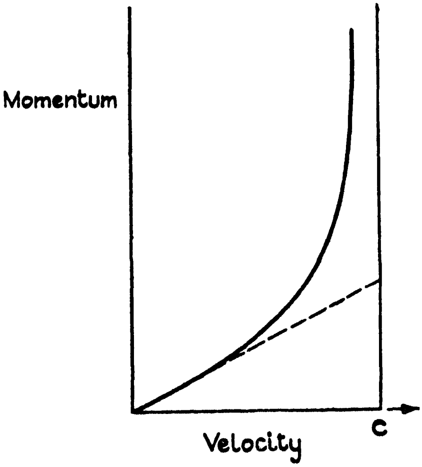


Fig. 34. Momentum and velocity.

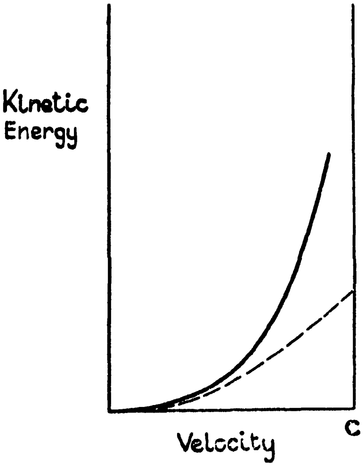


Fig. 35. Kinetic energy and velocity.

express the dynamical law in the form that force equals mass times acceleration. The force must be taken as the rate of change of the momentum, i.e. the rate of change of mass times velocity. These two are not the same when the mass also changes in the course of the motion.

This result says that as the speed of an object increases towards that of light, it becomes harder and harder to accelerate it. In fact, to make it reach the velocity of light in any finite time an infinitely strong force would be required. It is therefore evident that it is impossible for any material object ever to reach the velocity of light, let alone exceed it. This result is satisfactory for the consistency of our views. If it were possible to apply our considerations about the relations between different observers to a man on a train travelling with more than the velocity of light we would immediately become involved in paradoxes.

As regards the kinetic energy, the argument which I have sketched indicates that $E_{kin} = \frac{mc^2}{\sqrt{1 - \frac{u^2}{c^2}}} - mc^2$. This relationship is shown graphically in Figure 35, where again the full line gives the relativistic relation between kinetic energy and velocity, and the broken line the relation $E_{kin} = \frac{1}{2}mu^2$. As before, the Newtonian law is adequate at low speeds, but at high speed the energy of the moving object is much greater than it would be according to the old mechanics, and a body moving actually with the velocity of light would have to have an infinite amount of energy.

Using the definition of the variable mass, M, we may express our result for the kinetic energy as $E_{kin} = (M - m)c^2$ or, as the mass of the object increases the kinetic energy increases. The amount is exactly the increase of the mass times the square of the light velocity. This suggests a close relationship between mass and energy, though up to now only the kinetic energy entered into this. One can easily prove that also energy of any other form must lead to an increase in mass by precisely the same amount. To see this it is only necessary to consider the collision, not of two elastic objects, but of two completely inelastic objects, say two lumps of clay, of the same mass. If these collide with equal and opposite velocities they will just stick together. As the total momentum is zero the fused lump has no momentum and therefore is standing still. We may look at this process from the point of view of a moving observer and again require that momentum be conserved. If we work out the apparent velocities for the new observer we find that momentum would not be conserved if the fused lump had simply the combined rest mass of the separate lumps, but only if its rest mass had increased precisely by the amount of energy which had been converted into heat, divided by c^2. Hence the relationship between mass and energy includes also heat energy, and since we may arrange for the heat to cause some chemical reaction or to generate electricity by some internal mechanism, the same must apply to energy in any other form.

Hence we find the equivalence of mass and energy in the form that for any change of energy from E_1 to E_2, the mass will also change from M_1 to M_2, such that $E_2 - E_1 = (M_2 - M_1) c^2$.

If this is true for all changes in mass it is natural to speculate whether the rest mass m of a body is not also equivalent to some form of energy. Now in all our discussions so far energy has always appeared only through the law of conservation of energy, which compares the energy before and after some process has taken place. Therefore nothing is changed by adding a constant to all energies, provided this constant amount always appears on both sides of the balance. As long as objects do not disappear, or appear from nowhere, it is immaterial whether or not we include in the energy balance an amount mc^2 as their "rest energy", and it is convenient to do so since we can then generally state that all mass is equivalent to energy and vice versa. But this general statement implies more than convenience, because we shall see later that it is possible to create electrons and other material particles, and that the energy required for this is indeed given by mc^2.

APPLICATIONS AND CONFIRMATION

From reading this chapter so far the reader will have got the impression that we have built an elaborate structure of reasoning on one experimental fact, namely the negative result of the Michelson experiment. However reliable the experiment, and however attractive the general principle of the independence of all laws of physics of the state of motion of the observer, one would not have accepted such far-reaching conclusions without a great deal of further support. In fact other physicists have claimed from time to time that on repeating the Michelson experiment they did find a positive answer, but in the meantime other evidence for the theory of relativity had become so strong that we would have no cause to change our views if some fundamental flaw was discovered in Michelson's reasoning.

Such support comes from practically any observation of the behaviour of particles moving at high velocity. Historically the earliest observations made were on electrons, since, being the lightest particles, these can most easily be accelerated to velocities near those of light. For example, an electron reaches a speed of 0.8*c* after passing through an electric field with a potential difference of about 300,000 volts, which is now commonplace in any modern laboratory. At this speed its kinetic energy is about twice as large as it would be in non-relativistic mechanics and the difference in its behaviour very marked.

Earlier experiments carried out with much lower voltages, and in which the deflection of electrons in electric and magnetic fields was measured accurately, showed up the variable mass and therefore confirmed relativistic mechanics. Perhaps the most striking demonstration of relativistic behaviour results from a collision of a fast particle with one at rest. Supposing an electron has been accelerated somehow to a velocity near that of light and in passing through matter collides with a stationary electron. Since the two particles have the same mass and since usually in such collisions no mechanical energy is converted into any other form, our discussion of Chapter 1 of elastic collisions is applicable. We should conclude therefore that after the collisions the two electrons would move in directions at right angles to each other. On the other hand, relativistic mechanics says that both electrons should still travel forward, their lines of motion making only a small angle with each other, this angle becoming less and less as the velocity of the incident electron approaches that of light. We shall later discuss modern techniques which have made it possible to photograph the track of a single electron. Figure 36 shows a sketch of such a photograph, in which a fast electron is seen to be suddenly deflected from its path, and the track of a second electron originates from this point. All tracks are curved, since a magnetic field was applied in order to determine the speed of the electrons. After the collision the tracks form a very small angle with each other. This picture is a sketch of a photograph obtained by Dr. F. C. Champion.

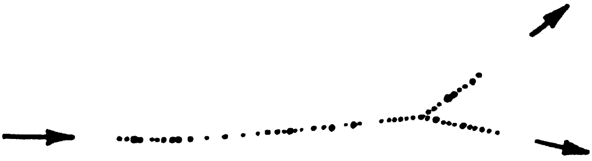


Fig. 36. Track of a fast electron colliding with a nearly stationary one.

Another drastic example is provided by modern particle accelerators, whose principle and purpose we shall discuss later. In these we have particles moving either along a straight line or (under the influence of a magnetic field) in a circle, and electric impulses are applied to them at suitable times. The performance of such machines depends therefore on knowing the speed of the particles so that the electric impulse can be arranged to be applied just as a particle passes a suitable point. Now there exist, for example, machines in which electrons can be given an energy of 300 MeV. By an MeV or million electron volts, we mean the energy which an electron would acquire on passing through an electric potential difference of one million volts. The rest energy of an electron is about $\frac{1}{2}$ MeV and therefore we are dealing with electrons whose energy is about 600 times its rest energy. A little arithmetic shows that this makes their velocity less than that of light only by a little over one part in a million. Without relativity, electrons of such energy should move with a speed of 35 times that of light, and it is evidently impossible to mistake one for the other.

These are only a few examples of the many tests to which relativistic mechanics has been subjected in recent years and the combined weight of these tests is sufficient to give us as much confidence in the relativistic laws, for fast-moving particles, as we have in the validity of Newton's laws in the domain of small velocities.

One particular consequence of relativistic mechanics is worth noting since we shall want to refer to it in the next chapter. If we consider the expressions for momentum and energy which have been given before, a little algebra shows that we can express them in the form of a direct relation between the two quantities, namely $E^2 = c^2p^2 + m^2c^4$, if E stands for the total energy of a moving particle, including its rest energy mc^2. If for a particle of a certain momentum p the rest mass m is so small that mc is negligible by comparison with p, this equation leads in a good approximation to $E^2 = c^2p^2$ or $E = cp$. We may consider the extreme case of a particle with no rest mass at all. For this the last equality is exact and its energy equals its momentum times the velocity of light. Such an object would always move with light velocity, whatever its momentum. The first expression for the energy $E = \dfrac{mc^2}{\sqrt{1 - \dfrac{u^2}{c^2}}}$ fails because both the rest mass m and the square root are zero, and zero divided by zero has no meaning. But the relation between energy and momentum remains valid even in this exceptional case. We shall meet in the next chapter the importance of this result for the properties of light.

PRINCIPLE OF EQUIVALENCE

The ground which we have covered in this chapter belongs to the so-called "special" relativity. Before we leave the subject of relativity we should mention briefly the wider subject of "general" relativity, though we shall not be able to do it justice.

Up to now we have been careful to compare only the points of view of observers moving relatively to each other with uniform speed. General relativity is concerned with the laws of physics as they might appear to an observer subject to acceleration. In this case it is clear that nature will look different to the moving observer. A passenger on a train which accelerates or brakes violently will notice the effect of this acceleration without question. What is the nature of his sensation? We have discussed this already in Chapter 1 and found there that this was a manifestation of Newton's law that mass times acceleration equals the force. In order to keep us moving with the accelerating train the seat or floor of the coach must exert a force on us. Without this force all objects would tend to keep moving uniformly and therefore, relatively to the train, acquire an acceleration in the backward direction. This is the same as if the train was standing still, but there was a field of force like gravity pulling us backwards. In fact, in a steadily accelerating train one easily forms the illusion that the coach is sloping and that our weight is therefore pulling us back.

Any acceleration produces the same effect as gravity. Imagine that we were in a closed box like the cage of a lift, which was placed somewhere in free space a long way from the earth or other bodies, so that there was no gravity. Then all objects in this box would float around freely. If we pushed our feet only slightly against the ground this would project us upwards until we hit the ceiling. This state of affairs has by now become familiar from books and films about space travel. Now suppose we suddenly felt our weight again and found objects falling to the floor. We could place two interpretations on this: either we could say that our box was really the cage of a lift and that it had started accelerating upwards, or we could say that the box was still stationary (or moving with uniform velocity), but that we were now near the earth, or near some planet, and there was gravity acting on us.

We could not settle the argument even if we had a small window in the ceiling and saw through it that there was a cable attached to the top of our cage, and that this cable was evidently in tension. This would still be expected on either view, either to produce the acceleration, or to keep our cage suspended against the force of gravity.

Conversely, if we were travelling in an ordinary lift and the supporting cable broke, so that the cage with all its contents was falling freely, we would have the sensation of the absence of weight, since the force of gravity would accelerate us at just the same rate as the cage, so that no force was required to keep our position relatively to the cage. We would therefore not know whether in fact our cage was falling or whether the attractive force of the earth had suddenly ceased (this doubt would of course last only until we hit the bottom of the shaft).

This is one of the few places in which the usually so reliable imagination of Jules Verne led him into a serious error. In his *Journey to the Moon* the passengers experience the sensation of weightlessness only at the place where the gravity of the moon and that of the earth just compensate each other. They should have experienced the same sensation all the way, as soon as the propelling mechanism had ceased to act, and their space ship had ceased to be subject to the friction of the earth's atmosphere, and was therefore moving freely in the earth's field.

We conclude that accelerated motion does not leave the laws of physics unchanged, but produces the same effect as a gravitational field. This "principle of equivalence", which is an essential feature of the general theory of relativity, depends essentially on the fact, already noted in Chapter 1, that the weights of two objects are exactly proportional to their masses, so that gravity gives them the same acceleration. The general theory of relativity is based on the view that an acceleration of all the landmarks to which we refer our observations is not merely in practice indistinguishable from the effect of a gravitational field, but that the two are essentially one and the same thing and cannot be distinguished in principle.

This does not mean to say that we could set up our laboratory in such a way that we could get rid of the whole gravitational attraction of the earth, for example, because this would require distributing observers in free falling boxes all around the earth, and of course they would soon hit the earth, unless they had previously collided with each other. We can only in a small region of space and for a short time look at the motion of objects from the point of view of an observer for whom there exists no gravity.

GENERAL RELATIVITY

If we tried to set up our description of nature in this way, doing our mechanics in a freely falling box we would come to the conclusion that space has curious properties. For example, suppose that our box arrives with high speed from large distances and the speed is so great that the attraction of the earth would not be sufficient to make it collide with the earth, but only deflect it through a small angle. In Figure 37 *E* shows the earth and the line marked *B* might be the path of our box. Suppose also that close to us a second box *C* with another observer was travelling parallel with us and at the same speed. Until we get near the earth, *C* would be at a constant distance from us, but being closer to the earth his box would be deflected more, as shown in the figure, and so afterwards we would be travelling in different directions and would be separating. We are both travelling freely and therefore experience no sensation of acceleration. On the ideas of general relativity we should therefore both be entitled to believe ourselves at rest, but we should then find the distance between us suddenly starting to increase without either of us apparently having started to move.

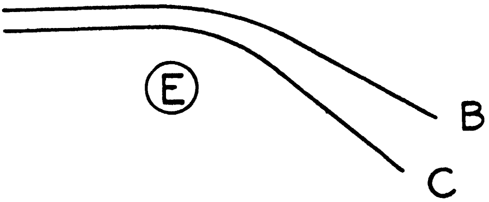


Fig. 37. Two observers passing the earth.

In other words, without answering the question about rest or motion and choosing as landmarks any objects which are allowed to move freely, we can detect the presence of a gravitational field and therefore the presence somewhere in the neighbourhood of heavy matter, by the way distances between different landmarks behave. This is the beginning of a description of a gravitational field in terms of the properties of space and time. It would require too much mathematics to pursue this description in greater detail, but I shall mention two conclusions which follow from it.

One is that in a gravitational field light does not travel in a straight line, but is deflected. This in fact could have been deduced already from the fact, proved in special relativity, that mass and energy are equivalent. A light ray carries energy and therefore mass, and if all mass is attracted by a heavy body like the sun or the earth, this should also hold for a light ray. However, the amount of deflection is not correctly given by this simple argument because the force of gravity on a moving object depends on its speed. This is unimportant for the slow speeds normally encountered in practice, but makes some difference when the speed is that of light. When light passes very close to the sun it should therefore be deflected by an amount which can be calculated, and therefore a star which we happen to see near the edge of the sun should appear slightly displaced from its normal position. It is of course not usually possible to see a star when it is close to the sun, but an opportunity for this arises during a total eclipse of the sun, when the sun's disc is covered by the moon so that its light does not make the stars invisible. Even then very accurate observation is needed to detect the displacement, since the expected amount is less then two seconds of arc, i.e. about the size of a penny seen from a distance of two miles. Nevertheless precise astronomical observations have made it possible to detect such displacements, and while astronomers still use convenient solar eclipses to improve the accuracy of the result, it is certain that the displacement exists and is of roughly the amount predicted by general relativity.

Another important conclusion is that the field of gravity should affect the time scale, or more precisely that a clock kept near a very massive star where there is a strong negative (i.e. attractive) gravitational potential, should appear to us to be slow. This, too, can be tested since the light from a hot star contains radiation of very distinct colours, i.e. of certain definite frequencies, which we know are due to the tendency of some atoms in the star's atmosphere to set up electric oscillations with a definite frequency, much like a miniature radio transmitter. Such atoms therefore may be regarded as standard clocks, in that the period of vibration is a natural characteristic of the atom and not dependent on external circumstances. Now it is known that from certain very massive stars these characteristic rays do not have the same colours as those from lighter stars, but their colour is more towards the red.

This agrees with the predictions of general relativity. We must assume that from the point of view of an observer on the surface of the star, the atomic "clock" would be correct, but owing to the dependence of the time scale on gravitational potential, to us watching from a distance, the clocks would appear to be slow and the frequency of the light smaller, and the colour of the light more in the red.

This influence of the gravitational potential on the behaviour of clocks is interesting also because of its connection with a paradox which illustrates the close connection between the ideas of special and general relativity.

It is sometimes objected that special relativity is inconsistent in its prediction of time dilation. Suppose two observers pass each other at high speed and make sure that when they pass their clocks show the same time. Now if we regard ourselves as standing still we would conclude that the clock of the other observer must be slow. He in turn would say that our clock should be slow because we were moving at high speed. Admittedly this cannot be checked as long as we are a long way from each other, since we have seen that we cannot unambiguously compare the times of events taking place in distant places. However, supposing later we meet the other observer again, then we can compare our clocks and we shall know whose clock is slow with respect to the other and therefore who has been moving.

Within special relativity the correct answer to this objection is that as long as the two observers keep moving at uniform speed relatively to each other, they can never meet again. If they are to meet again at least one of them must either reverse his motion, or change direction, and therefore be subject to acceleration, so that he is no longer entitled to assume that he has been at rest all the time.

General relativity provides a more complete answer. Suppose for the sake of argument we have been moving without acceleration, whereas the other observer after travelling some distance away stopped and returned. In general relativity he would still be entitled to assume that he had been standing still all the time, but that there was a gravitational field present. This would account for the forces that he felt acting on him (which we would ascribe to his acceleration). Now gravitational fields, as we have seen, affect the clocks, and if this is applied in detail to the situation under discussion the answer is that again on both views the readings of the two clocks would compare in the same way.

We shall leave the interesting subject of general relativity after this very incomplete discussion. A fuller study of it would certainly belong to a complete account of the laws of nature, but our aim here is to present the laws that have a particular bearing on the structure of matter. General relativity has important consequences for large-scale problems, including the structure of the universe. It is likely that in the future some relationship may emerge which links these ideas also with the characteristics of small-scale, i.e. atomic problems, but this is as yet unknown.

7

Quanta, Waves and Particles

THE QUANTUM HYPOTHESIS. PHOTONS

CHAPTERS 4 and 5 gave an outline of the elementary facts about atoms. We had seen that matter was built of atoms, whose mass and size could be determined, that their motion accounted for heat, their mutual forces for the rigidity of solid matter, their impact on walls for the pressure of a gas. We have seen that each atom consists of a heavy positively charged nucleus at the centre surrounded by light, negatively charged electrons. But while, therefore, this picture in many ways accounted for the properties of matter, again and again we met contradictions, the most outstanding one of them being the fact that atoms were always of a definite size and did not collapse under the attraction of the positive nucleus on the negative electron.

We shall not here go over the whole list of such contradictions, both because it is too long and also because some of them will have to be raised again later to show how they are resolved by the principles of the new mechanics. However, the examples already explained should suffice to show how it became clear that the laws of nature outlined up to this point are not adequate to describe the behaviour of atoms.

What is required is evidently something which determines the scale of atomic phenomena. By this we mean the following: if we look at the problem of an electron moving around a positive charge under the influence of forces obeying the inverse-square law, then the particle can move, for example, in a circle of any radius we might choose, provided its speed is adjusted so that the centrifugal force just balances the attractive electric force. We have discussed such circular motion in detail in Chapter 1. The same is true in the solar system. There is nothing in the laws of mechanics which says that the size of the orbit in which the earth revolves round the sun should be just what it is. If it were a hundred times smaller or a hundred times larger, this would be perfectly possible mechanically, only the speed of the earth would have to be ten times larger or ten times smaller, respectively. In the case of the earth, the actual motion is the result of a much earlier history. Some very long time ago, the earth and the sun were part of the same mass, probably of very hot gas, from which matter collected into large lumps. It was the speed of rotation of that gas mass and the particular way in which matter condensed, that gave rise to the actual positions of the earth and the sun. But in the case of the atom, we know that its size is not an accident of its previous history. We can, in many ways, throw the electron out of a hydrogen atom and allow the remaining hydrogen nucleus or "proton" to recapture an electron; when this has settled down again to a steady state, the hydrogen atom is indistinguishable from any other.

To explain this became possible only through the "quantum of action". This was introduced by Planck by means of a very bold hypothesis. Planck was studying the black-body radiation which, as we have seen in Chapter 5, also presented features which conflicted with the laws of "classical" physics; see, for example, Figure 28. Planck's idea could account for the amount of radiation from a hot body and its composition in colours.

He postulated that light carried its energy in definite amounts or "quanta", so that the energy of the wave always had to be one, or two, or any other whole number of such quanta. The energy of each quantum of any given colour had to be the same, and proportional to the frequency of the light. For light of frequency ν (i.e. for a wave in which the electric field at any point in space goes through ν cycles a second) each quantum thus has the energy $h\nu$, where h is a new constant, called "Planck's constant" or the "quantum of action" It is approximately 6.5×10^{-27} in metric units.

The important point is that, when an atom emits radiation of a certain frequency (or colour) it can, according to this hypothesis, produce one or more such quanta, but never a fraction of one. Similarly an atom exposed to light can absorb only one or more, but not a fraction of a quantum.

This rule sounds very unreasonable and it appeared so to Planck. He attempted to use it in various modified forms, which would make it appear less extreme, and it was only later developments which demonstrated that nothing less than the full rule would really do. In some way this appeared to be going back to the old emission theory of light. If light really consisted of little particles, which were ejected by the source of light, and which travelled until they reached the eye, or a photographic plate, or some other instrument which responded to light, then it would not be unreasonable to believe that the particles belonging to light of a particular colour, i.e. of a particular frequency, should always carry a fixed energy. The light beam of that colour would then always contain a whole number of such particles. But how should we then explain the evidence showing light to consist of electromagnetic waves? This apparent contradiction will have to remain unresolved until later. One of the most important lessons of modern physics which we shall have to learn, is how it is possible to reconcile these apparent contradictory statements, i.e. that light consists of waves and that light consists of distinct "quanta" of determined energy.

For the moment we just note the fact that if this bold hypothesis is made, then the curve of Figure 28, which indicates the intensity of black-body radiation of a given colour at any temperature, follows naturally from it. To demonstrate this in detail would require some mathematics, but it is possible to see at least that the hypothesis changes the result in the right direction. Consider a frequency ν and take a temperature T such that kT is considerably less than $h\nu$. Then the argument of Chapter 5 would have led to the result that the energy contained in the wave should be equal to kT and therefore be a small fraction of the energy of one quantum $h\nu$. Now according to the hypothesis of Planck this is not possible. The energy must either be nothing or a whole quantum (or more). We have seen in Chapter 5 the random character of heat motion, so the energy available to a light wave will sometimes be larger and sometimes smaller. It might well sometimes amount to one quantum and sometimes to none. But the chance that it amounts to a whole quantum comes out to be very small, because this can only happen if by chance one or more of the atoms in the hot body which produces the radiation are in exceptionally violent motion, i.e. if the heat motion happens to be very much more violent than it is on the average.

To use a crude analogy, suppose I run a business in which the average amount of cash in hand is £100. Then on the average my cash box will contain £100 at the end of each day. But if I have instead a safe with the rule that I am allowed to deposit in the safe only money in "quanta" of £1000 at a time, then most of the time the safe will be empty. Only at the end of a very exceptional day, when there have been very many customers and no bills to pay at all, can I make a deposit in the safe. This will happen much less frequently than every tenth day and therefore the average contents of the safe will be much less than £100.

As long as there was no other evidence to support Planck's hypothesis, one would have looked at it with considerable reserve, because it is often possible in physics to explain one particular set of observations by means of a hypothesis introduced for that purpose, and this need not have any real significance. But there were immediately two other observations which had not previously been explained and which fitted in very well with the idea of Planck. The first of these was the photo-electric effect.

If light falls on a metal surface in an evacuated glass tube, across which there is an electric field, an electric current may flow across the tube, carrying negative charge from the illuminated plate to another. It was demonstrated that this current consists in fact of electrons which are removed from the metal plate by the action of light. This, in itself, is perfectly natural, if one remembers that light consists of electromagnetic waves. A metal is known to contain electrons and on them the electric field of the wave should exert a force. If this force was strong enough, it might be sufficient to pull electrons out of the metal. But on this view, one would have expected that this so-called "photo-electric effect" would require a strong light. If the light was very weak, then the electric field associated with it would be very weak and a weak electric force would not be sufficient to overcome the forces of attraction which normally keep the electron within the metal. We would also expect that if the light intensity were increased, then the velocity with which the electrons are thrown out would increase because they are pulled out by a stronger force. Finally, the dependence of this effect on the light frequency might not be simple, but on the whole, we would expect light of higher frequency to be less effective. If the force reverses its direction after too short a period, it will not have a chance to pull an electron out of the metal before it starts again to push it back in the opposite direction.

But observation showed quite different results. It was shown, first of all, that for a given colour of light, i.e. a given light frequency, the electrons range in speed up to a maximum which was the same for weak or strong light. If the light was weaker, fewer electrons come out. The number of electrons varies directly as the light intensity, i.e. if the intensity is reduced ten times, one obtains ten times fewer electrons, but there is no "threshold", i.e. no definite minimum intensity below which there were no electrons. On the other hand, there is a very definite relation with the light frequency, i.e. with the colour of the light. The kinetic energy E of the fastest electrons ejected from the metal can be expressed as $E = h\nu - W$, where h is again Planck's constant and W an amount of energy which varies from metal to metal, but which for each metal is the same regardless of the intensity or frequency of the light which is used. If $h\nu$ is less than W, there is no photo-electric effect since the kinetic energy cannot be negative.

Einstein pointed out that this was exactly what one should expect on the basis of Planck's hypothesis. If light consists of definite quanta, then as such a quantum hits the metal surface, it may be absorbed, i.e. disappear from the light beam so that its energy becomes available. If this amount of energy $h\nu$ is passed on to an electron, the electron may be liberated from the metal and obtain some speed in addition. If W is the amount of energy necessary to remove an electron from the metal, then clearly the amount $h\nu - W$ would be left over for its kinetic energy.

If we increase the light intensity, we change merely the number of quanta which strike the metal surface during every second and therefore the number of electrons produced.

This result therefore indicates that there is something constructive in Planck's idea, since it leads automatically to the correct behaviour of electrons in the photo-electric effect. It is particularly important that in this relation the same constant h appears which had already been determined from black-body radiation.

A second experiment which demonstrates in a convincing way that there is something real about the idea of light quanta is the Compton effect. This consists in the discovery that when X-rays or gamma-rays (i.e. electromagnetic radiation of a wavelength similar to atomic dimensions or shorter) pass through matter, one observes that some of the radiation is scattered, i.e. deflected into all possible directions. This in itself is again what one would expect from the older ideas. The electromagnetic wave exerts a force on the electrons in some of the atoms, causing them to oscillate rapidly, and such rapidly oscillating electric charges act as small transmitters for electromagnetic waves and therefore produce radiation in all directions. However, we should then expect that the frequency of the scattered radiation would be the same as that of the original beam. Compton discovered that the frequency of the scattered radiation was always less than that of the radiation which came in. This decrease in frequency is greater, the greater the frequency of the incident radiation and the greater the angle at which the scattered radiation is observed.

Compton gave a theory of this result based on Planck's idea. The scattering process should then be regarded as the deflection of a light quantum into a different direction. Now a change of direction must also change the momentum of the light quantum, similarly to the case of a mechanical object (such as a billiard ball) which cannot be deflected from its straight path without a force and therefore without some other object taking up the change of momentum. If a light quantum has energy $h\nu$ its momentum must be $\frac{h\nu}{c}$. This follows from the result we had derived at the end of Chapter 6. A light quantum evidently must move with the velocity of light and cannot have any rest mass. Therefore the result obtained in Chapter 6 is applicable, according to which the momentum of such a body is its energy divided by c. The same result can also be deduced from wave theory. If we consider an electromagnetic wave travelling in a given direction and carrying a certain amount of energy E, we can work out the amount of momentum associated with this and we get the result $\frac{E}{c}$.

If we now assume that the light quantum in the Compton effect has been interacting with an electron inside some atom near which it passed, then we must expect the electron to recoil, i.e. to change its momentum by such an amount as to balance the momentum change of the light quantum. To discuss this quantitatively, let us assume that the light quantum has been deflected by a right angle. Then we have the situation sketched in Figure 38, in which the arrow marked $\frac{h\nu}{c}$ represents the momentum of the light quantum before the collision, the arrow marked $\frac{h\nu'}{c}$ that after the collision, and the broken, arrow marked p the momentum of the electron after the collision. These three arrows must form a triangle, as shown, to satisfy the law of conservation of momentum. Here we have taken the angle between the first two arrows to be a right angle and we have then from the rule of Pythagoras that $p^2 = \left(\frac{h\nu}{c}\right)^2 + \left(\frac{h\nu'}{c}\right)^2$. In addition, energy must be conserved and therefore $h\nu - h\nu'$ must equal the kinetic energy of the electron which, according to Chapter 6, is the difference between the total energy $\sqrt{(m^2c^4 + c^2p^2)}$ and the rest energy mc^2. From these two equations some straightforward algebra leads to the result that $\frac{1}{\nu'} - \frac{1}{\nu} = \frac{h}{mc^2}$ or, introducing the wavelength $\lambda = \frac{c}{\nu}$, we

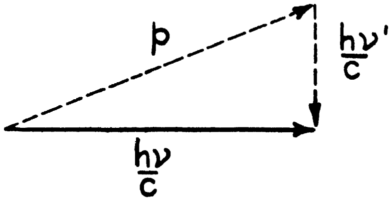


Fig. 38. Momentum balance in Compton effect.

have $\lambda' - \lambda = \frac{h}{mc}$. In other words the wavelength of X-rays scattered through a right angle increases by $\frac{h}{mc}$ where m is the rest mass of the electron.

The problem can be solved equally easily for a different angle of scattering. This result is in excellent agreement with observation, and we see that the idea of a light quantum now takes an even more tangible form. We have to think of it as a particle carrying not merely energy but momentum, and capable of colliding with electrons in just the same way as one electron might collide with another. It is possible, also, to observe not merely the scattered radiation, but also the electrons which, according to our explanation, should be projected whenever X-rays or gamma-rays are scattered, and these again are found. The speed of the electrons projected into any given direction again confirms exactly the prediction of the theory.

The quanta of light energy which, are given a very tangible form by the last results, are often also called "photons".

We must, however, recognize that these new ideas form a radical departure from earlier ideas about light and other radiations. While their success in explaining certain observations is striking, they appear, as has already been stressed, to be in complete contradiction with other known facts concerning the wave nature of light, in particular with the interference and diffraction of light waves.

ELECTRON DIFFRACTION. STANDING WAVES

This difficulty is not peculiar to the case of light. Whereas light, which we had proved to consist of waves, seemed suddenly to acquire, in some respects, the characteristics of beams of particles, it was later found that also beams of particles, such as electrons, equally have the characteristics of waves. The most striking demonstration of this consists in the diffraction of electrons, which was actually predicted by quantum theory before it was found experimentally. If a beam of fast electrons of uniform speed strikes the surface of a crystalline substance, one observes electrons reflected only in certain directions. The angles of intense reflection are exactly the same as for the diffraction from the same crystal of X-rays of a particular wavelength as discussed in Chapter 4. In practice the experiment is somewhat more difficult than with X-rays, since beams of electrons are more easily deflected from their straight path by accidental electric charges or by matter through which they pass, so that for clean conditions the experiment must be performed in a vacuum and with other important precautions.

Now the diffraction of light and of X-rays had been the most important piece of evidence to convince us that these consisted of waves; if we find that electrons can be diffracted as well then it follows also that electrons are waves. The wavelength which can explain the diffraction pattern for an electron beam of velocity v turns out to be $\frac{h}{p}$, where h is again Planck's constant and p the momentum of the electron. When the velocity v of the electron is much less than the velocity of light we may also express the momentum as mass times velocity and the wavelength then is $\frac{h}{mv}$.

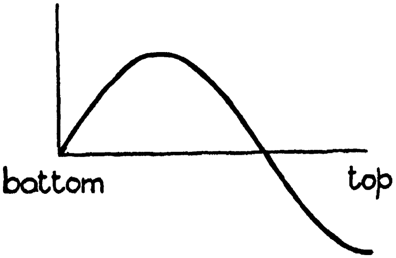
The statement that the wavelength is Planck's constant divided by the momentum is true not only of electrons only but also of light. Indeed, we had seen before that the momentum of a light quantum is $p = \frac{h\nu}{c}$, where ν is the frequency of the light wave to which it belongs; in other words, the wavelength λ, which equals $\frac{c}{\nu}$, is again equal to $\frac{h}{p}$.

This analogy between electrons and photons was, in fact, the starting point of the work of de Broglie, who suggested before the discovery of electron diffraction that the electron might have wave properties. By comparing electrons with photons, he predicted that the electron wave should have a wavelength $\frac{h}{p}$.

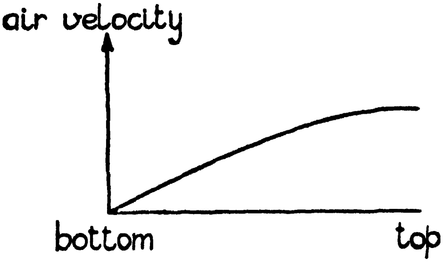
All this does not answer the question how it is possible that light can show all the characteristics of waves and at the same time behave as if it consisted of separate particles. We can learn from electron diffraction merely that the same paradox applies also to electrons. In fact, this holds for particles or objects of any other kind. We must therefore not look for an explanation which is tied particularly to the nature of light. This difficulty brings us up against one of the most important, but also one of the most difficult, parts of modern physics. Before we face it, it is useful to consider some further consequences of the wave picture.

For the moment, let us therefore accept the fact that electrons are associated with waves, though we must remember that many questions about this as yet remain to be answered.

One characteristic of waves which will be of the greatest importance is that a wave confined to a limited space can only have certain selected wavelengths which fit into the available space. Examples of this type of behaviour are probably most familiar from sound waves. Any musical instrument produces notes of a certain pitch, i.e. waves of a certain frequency and wavelength. Take as a simple case a tube, closed at the bottom end and open at the top; this permits the air in it to move in a wave motion in which it stands still at the bottom end and has its greatest velocity at the top end, giving therefore a "node" at the bottom, and a wave crest at the top end. In other words, the longest possible wave is four times the length of the tube, so that the tube represents one quarter wave. This gives the lowest or "fundamental" note of such a tube, and it is also the one that is easiest to produce. It is illustrated in Figure 39. However, a motion is also possible in which there is an intermediate trough in the wave so that the tube corresponds to three-quarters of a wavelength. Similarly it may be five-quarters, and so on. (In practice, these "harmonics" are not always easily produced by themselves, but in



First harmonic



Fundamental note

Fig. 39. Air movement in a tube.

most wind instruments, for example in the recorder, they are used for the higher notes, i.e. the shorter wavelengths.) In other words, in such cases, the waves inside the instrument belong to one of a sequence of sharply defined wavelengths.

Similarly, in the wave motion of a violin string, the wavelength must be such that both the bottom end of the string and the point where the finger of the violinist rests represent nodes, and possible wavelengths are such that either one half, one, one and a half . . . waves fit in between those two points.

The same applies to waves of other kinds. A typical example of water waves arises in a bath when the wavelength can be for example twice the length of the bath, so that the water rises at one end and falls at the other. If a person sitting in a bath tub inadvertently moves his body back and forth a few times with the frequency corresponding to just that wavelength, he may set this particular wave going so strongly that the water rises over the edge of the bath tub at one end.

If electrons are waves, the same will be true for them. If we imagine an electron enclosed in a small "box" so that it will run back and forth between the walls of this box, then it should not be capable of performing this motion with just any velocity, i.e. any wavelength, but only with one of a sequence of sharply fixed wavelengths which will just fit into the box. Therefore the electron must move with one of a sequence of sharply defined velocities. Its kinetic energy must also equal one of a sequence of selected energies. The possible energies of an electron in such a case are called its "energy levels". According to the wave theory, therefore, an electron moving in a confined space has a sequence of sharply defined energy levels.

Now we have already remarked in Chapter 4 that we should expect an electron which moves back and forth in a confined space to give rise to electromagnetic radiation, i.e. to light, and in this way, as radiation carries away energy, the electron energy should diminish. On classical mechanics, the period of the light wave will equal the period of the electron in its box, i.e. the time which the electron takes to move across the box and back. If the electron loses energy, its motion will become slower and the time for its return journey longer. Hence in the course of its motion, the electron should produce radiation of all frequencies, from the highest frequency, which corresponds to the beginning of its motion, down to frequency zero. On the other hand, in the wave picture the electron is capable only of certain energies. If it starts, let us say, with an energy E_1, it must lose, if it loses any energy at all, at least the amount that will bring it to the next lower level, say E_2, and if any light is carried away, it must carry an amount of energy $E_1 - E_2$. Since light consists of quanta, the amount of energy which the electron has lost must equal a quantum of the radiation which has been produced. Therefore, $E_1 - E_2 = h\nu$, if ν is the frequency of the emitted light. In other words, the electron can produce only light of frequency $\nu = \frac{E_1 - E_2}{h}$. Now the electron finds itself at the energy level E_2, and it may further lose energy by going to the next lower level say, E_3, which can only produce light of the frequency $\frac{E_2 - E_3}{h}$. Alternatively, from the first level E_1 the electron could jump directly to E_3, thus producing light of the higher frequency $\frac{E_1 - E_3}{h}$.

The wave picture therefore leads to the result that radiation from such an electron will contain only sharply defined frequencies, each of which, when multiplied by h, represents the difference between two of its energy levels.

Conversely, the electron can gain energy by absorbing light which comes from an external light source, but again it will be capable only of absorbing light of the same frequencies.

These conclusions are completely in line with what had been known for a long time about the behaviour of atoms. Ever since the discovery by Fraunhofer of the sharp lines contained in the spectrum of sunlight, indicating that light of certain sharply defined frequencies was missing from the spectrum, physicists had studied the appearance of such lines when light was emitted or absorbed by gases. They had also found that the frequencies of these lines can be shown to be the differences of quantities which were characteristic for each atom and which we now know to be its energy levels, divided by h.

For an electron in a box a further important consequence is that its lowest energy level belongs to a finite wavelength, which will be about twice the width of the box (though this depends a little on its shape), and this greatest wavelength means a smaller velocity. Once the electron has the energy belonging to this least velocity it clearly cannot lose any more energy because it is not capable of having any less. Therefore it stops losing energy before it has stopped moving. This motion is called "zero-point" motion. It is clear that we have here a result which goes in the right direction to explain the fact that atoms do not go on losing energy until they collapse.

VARIABLE WAVELENGTH

Before we can see that all these conclusions apply to real atoms, we must get away from the rather artificial idea of a box, which had been introduced only to see the problem in its simplest form. In what sense is the motion of an electron, under the influence of the attractive force of a nucleus, similar to motion in a box? One difference between these problems is that in a box the electron moves always with the same speed, changing direction only when it bounces off a wall, whereas we know that in moving around an attractive centre, the speed of the electron changes, being greater when it is closer to the centre and less when it is further away. We had seen in Chapter 1 how this was related to the conservation of energy. The total energy E of the electron is made up of the potential and kinetic energy. As the electron approaches the nucleus, its potential energy is lowered (as a result of the attractive force) and therefore its kinetic energy must increase. Given the total energy E, we can therefore say what the velocity and hence the wavelength of the electron should be when it passes any particular point in space.

When we introduced potential energy in Chapter 1, we found it convenient to measure it in such a way that it is zero for a particle a long distance away from the attractive centre. Then close to the centre, the potential energy is negative; this means we must do work against the attractive force to remove the electron from the neighbourhood of the nucleus.

Hence the total energy E is made up of the negative potential energy V and positive kinetic energy, E_{kin}. The energy may be positive or negative, according to which of the two parts is larger. Take the case when it is negative. The orbit of the electron calculated by Newtonian mechanics is then a closed one, limited to a small neighbourhood of the nucleus, since the kinetic energy can never be negative, and therefore the electron cannot get to a place where V exceeds E. Negative energy therefore corresponds to a closed orbit like that of a planet around the sun. Hence, while in this case we are not concerned with any actual box, the electron is in fact confined to a narrow region. The only new features of the problem are firstly, that the size of the region accessible to the electron is not given once and for all as in the case of the box, but depends on its energy, and secondly, that we are not dealing with a case of constant wavelength but with a wavelength varying from place to place.

The precise method of describing waves of varying wavelength is through "differential equations". The particular equation appropriate to the wave representing an electron near some field of force is Schrödinger's equation. We shall not write it down here as such,

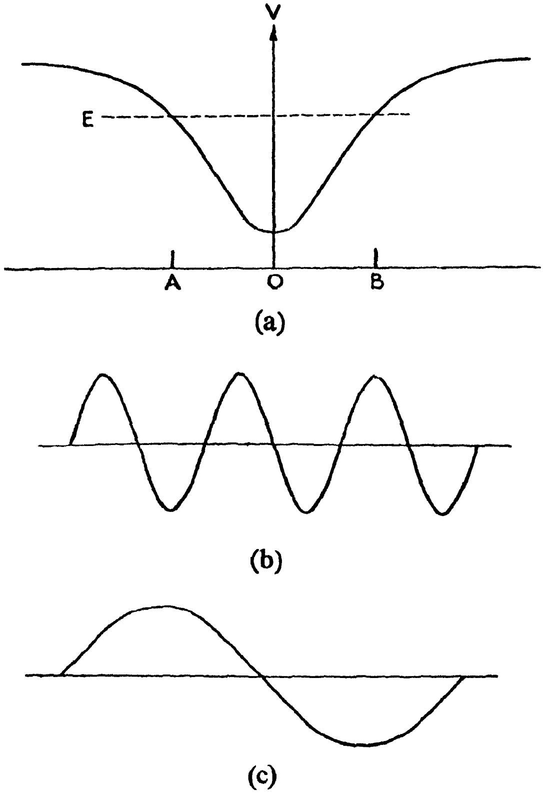


Fig. 40. Simple case of attractive force:

(a). Potential.
(b). Short wave.
(c). Long wave.

but we shall try to understand its contents and the consequences that follow from it.

The most important result is that, as for a particle enclosed in a box, the electron in an atom has a sequence of possible energy levels, and that the lowest of them still has a finite energy, with the wave extending a finite distance away from the nucleus.

To show how such a problem of variable wavelengths is dealt with, consider the simpler case of an electron moving only in a straight line under the influence of an attractive force which pulls it towards a fixed point *O* on that line. Figure 40a shows what the potential energy of such an electron might look like in a simple case, when plotted against the distance *x* from *O*. The potential energy is least at *O*, rising to either side because work would have to be done against the attraction to move the particle from *O* to anywhere else. Let the energy be *E* as shown in the broken line in the figure. In other words, *E* is greater than the potential energy at *O* but less than the potential energy very far away. If we regard this as a problem in classical mechanics, the electron could only move between the points marked *A* and *B* because outside of them the potential energy is greater than *E* and, since the kinetic energy is the difference $E - V$, it would have to be negative, which is impossible. At any point between *A* and *B*, we can find the kinetic energy $E - V$ as the height of the broken line above the full curve at this point. It is evident that the kinetic energy increases from *A* to *O* and then decreases again until it is zero at *B*. In the wave theory or "wave mechanics" we must therefore think of a wave not of constant wavelength but of a short wavelength near *O*, which becomes progressively longer as we go towards *A* or *B*.

Clearly we must extend our idea of wavelength a little. We can no longer think of the wavelength as being just the distance from one wave crest to the next one. The correct extension is obtained by noting that the profile of a short wave (Figure 40b) is, at every point, curved very strongly towards the horizontal axis, whereas that of a longer wave (Figure 40c) of the same elevation has much less curvature. A wave of infinite wavelength would, in fact, be represented by a straight line which has no curvature. We expect, therefore, that the curve representing the electron in a problem like that of Figure 40a should be very sharply curved near *O*; its curvature should become less and less towards either *A* or *B*; its curve should be just straight at *A* or *B*. Beyond these points, where in the classical motion the potential energy would exceed the total, we expect a curvature in the opposite direction, i.e. away from the base line.

We expect the wave, therefore, to look as in Figure 41a. From the centre towards *A* and *B*, the curvature gets less and less, and beyond the points *A* and *B* it is away from the axis. In this way we can have a curve which approaches the axis far away, so that the wave is negligible at large distances. In this case, therefore, the wave representing the electron extends somewhat beyond *A* and *B* to the region which, to an electron with the given energy, would

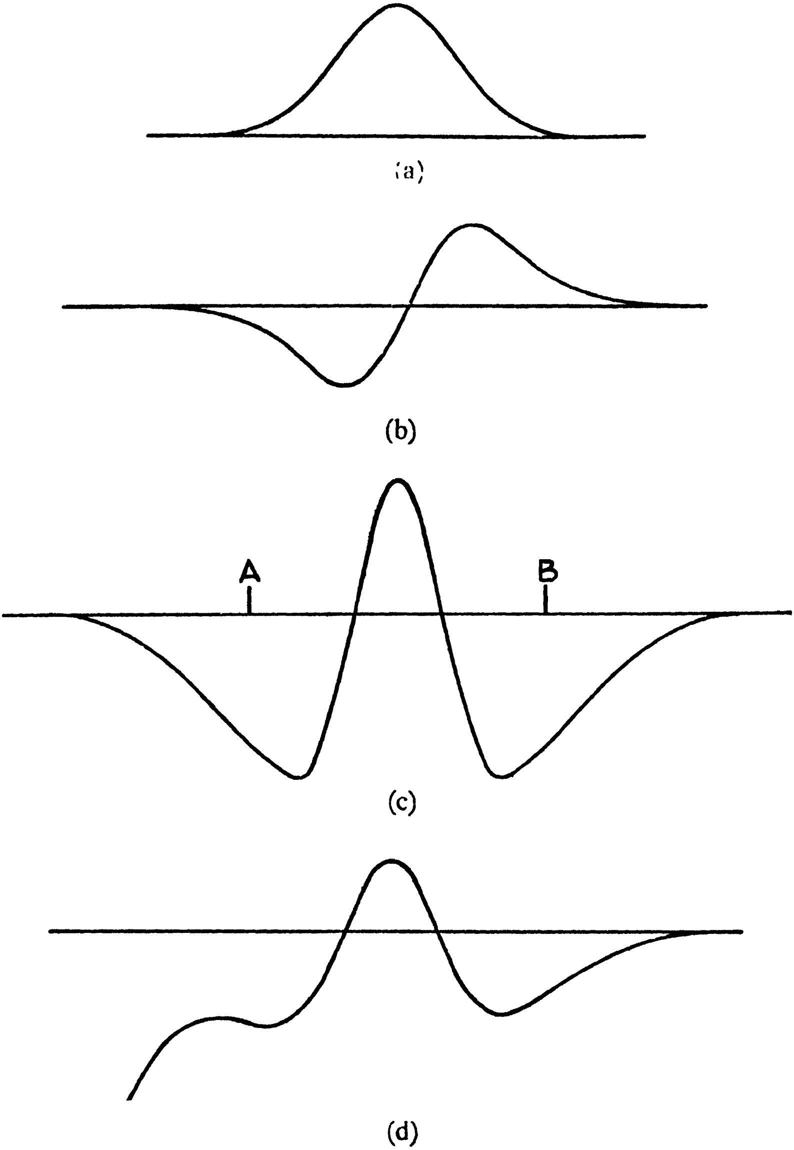


Fig. 41 (a to c). Possible waves in the potential of Fig. 40.
(d). An example of an impossible one.

not be accessible according to classical mechanics. This is a result to which we shall return.

Now imagine that in place of the energy E, we take one that is just slightly lower. Then the points A, B move a little closer together and in the whole region between them the kinetic energy is reduced; this means increased wavelength or reduced curvature of the wave profile. On the other hand, at large distances, the curvature away from the X-axis is increased. In that case it is impossible to obtain a curve looking like Figure 41a in that it approaches the x-axis at both sides. Instead, if we try to construct such a curve starting from the right and arranging for it to behave there like Figure 41c, we obtain something like 41d, in which, on the left, the curve runs away faster and faster from the x-axis, giving therefore a wave which would have infinite elevation a large distance away; this is not a reasonable solution to our problem.

In fact, a curve like that of 41d is what we must expect in general if we choose the energy at random; and only very special energy levels will correspond to curves like 41c in which we can obtain a reasonable behaviour at both ends. One can prove that the lowest of these energy levels corresponds to a wave without any nodes, i.e. a curve like 41a, which does not cross the axis. The next one has one node, the following one two, etc. Figure 41a-c shows a sequence of such functions which, in fact, were obtained by solving the Schrödinger equation for the potential shown in Figure 40a. The important point is that even the first of them, which corresponds to the lowest energy level, must, of necessity, have a region in which the profile is curved toward the axis. In this region, the energy E belonging to this wave must definitely be higher than the potential energy; in other words, the lowest energy level of this problem lies higher than the potential energy at the attractive centre O.

THE HYDROGEN ATOM

We see therefore that the electron wave in an attractive field of force behaves in a very similar way to that of an electron enclosed in a box, provided the energy of the electron is not sufficient for it to escape altogether from the attraction. The position is precisely similar in the real case of the attractive force due to a nucleus, for example to a hydrogen nucleus or "proton". Here, the electron wave covers a region of space surrounding the proton. To specify it completely, we should be able to show where the peaks and crests are anywhere in space, and this makes it impossible to state the wave in terms of simple diagrams as in Figure 41. But by analogy with the simpler case the following results are, at least, not surprising: for an electron moving near a hydrogen atom the possible energy levels are $E = -\frac{Rh}{n^2}$, where h is Planck's constant, R is the so-called Rydberg constant, which can also be expressed as $R = \frac{2\pi^2 me^4}{h^3}$ and n may be 1, 2, 3 . . . or any other whole number. No wave exists for any negative energy other than these levels.

For the "ground state", i.e. the lowest energy level $n = 1$, there exists only one wave, but there are four waves of different shape belonging to the second energy level with $n = 2$, and generally n^2 different waves to the nth level. We shall return to the question what distinguishes these different waves from each other.

For the ground state, $n = 1$, the electron wave is concentrated around the proton and falls off with distance. It is spread approximately over the distance $a = \frac{h^2}{4\pi me^2}$. This is approximately $\frac{1}{2} \times 10^{-8}$ cm., i.e. corresponds with the size of an atom as derived by the arguments of Chapter 4.

In this way, wave mechanics leads to the conclusion that there exists for an electron moving under the attraction of a proton a sequence of energy levels, and the electron can change its energy only by going from one to another of these levels; therefore it can receive or lose energy only in amounts which equal the difference between two of these levels. Coupled with what we know of the quantum nature of light, this shows therefore that light or other electromagnetic radiation sent out or absorbed by a hydrogen atom must always belong to frequencies which equal one of these differences, divided by Planck's constant h. This agrees exactly with observation. It has been known for a long time that most atoms, including hydrogen, give rise to a "line spectrum", i.e. that if the light produced by very hot hydrogen gas is passed through a spectroscope, which separates different frequencies from each other, there appear sharp lines. Conversely, if white light is passed through a space containing hydrogen atoms, the spectrum of the light shows dark lines in the same places, showing that light of these particular frequencies has been absorbed by the atoms. For hydrogen, the formula for the energy levels which has been quoted above gives exactly the observed position of the lines.

In explaining this situation, we have put the historical development, to some extent, in reverse. It had been known for a long time that the characteristic frequencies occurring in the spectrum of an atom could be obtained as differences between a sequence of numbers, and Niels Bohr combined this with Planck's theory of the quantum nature of light and saw that this indicated sharp energy levels of the atom. Bohr then developed his quantum rules, which allowed him to predict which of the many orbits that one could calculate from Newton's mechanics should be identified with the energy levels of the hydrogen atom. We shall not consider these quantum rules in detail, because they involve more subtle and more complicated arguments than the wave-mechanical picture which later superseded them. But we note that, from these quantum rules, Bohr was able to give the correct formula for the energy levels of the hydrogen atom. This success of the theory and a good many other predictions which agreed with observation left no doubt that Bohr's quantum conditions were essentially correct, but the position was not logically satisfactory because these conditions were superimposed on the old mechanical laws, simply selecting amongst the mechanical orbits those which conformed to the new conditions.

Thus, while some consequences of the old mechanics were retained, the new rules changed the nature of the mechanical problem completely. For example, according to the old mechanics, the energy of an electron should, under the influence of a light wave, change only gradually. According to Bohr's rules the energy could equal only one or the other of certain specified amounts, so that it seemed impossible to describe how the electron got from one orbit to the other. In spite of these difficulties, Bohr's ideas were applied to a large number of problems, and in many cases they gave the right answers, though to other problems they would give ambiguous or no answers. This situation remained until de Broglie pursued the analogy between light and electrons to the prediction that electrons should also be connected with waves, and he recognized that this picture would lead in some cases to standing waves and hence to selected frequencies or energies. This idea was then taken up and extended by Schrödinger, who gave a complete treatment of such waves and showed that they led to the same result as Bohr's quantum rules in all those cases in which the latter had been confirmed by observation. But they also provided definite answers in many cases to which Bohr's rules were not directly applicable.

This development linked up with ideas which Heisenberg had developed a little earlier, and which were based more closely on Bohr's rules, but contained an attempt to make them into more complete mechanical laws, and to make them applicable to problems for which they had not worked before. It was then shown that Heisenberg ideas were, in fact, the same as Schrödinger's in a different form and I have here presented the results in Schrödinger's form because they are then easier to explain in non-mathematical language.

Shortly afterwards the hypothesis of electron waves was directly confirmed when G. P. Thompson and, independently, Davison and Germer, discovered the diffraction of electrons, which we have taken as our starting-point in explaining our present knowledge.

PARTICLES AND WAVES

What we have learnt about electron waves and their behaviour in fields of force constitutes a set of definite laws embodied in Schrödinger's wave equation. For these waves this plays the same part as Maxwell's laws for electromagnetic waves or light. But there remains the paradox which had troubled us earlier in this chapter; electrons have sometimes the characteristic behaviour of waves, as in diffraction, or in the sharp resonances of a hydrogen atom, and sometimes the characteristics of distinct particles, shown by the fact that when we look for electrons we always see one whole electron, or more than one, or none, but never half an electron. This puts electrons on the same footing as light quanta, of which we had also found that they had a wave nature and a particle nature. What is the connection between those two aspects and how can we reconcile these two kinds of behaviour?

To understand this further, consider some process involving the propagation of light. We use here the term "light" for brevity, but it is immaterial whether we are thinking of actual visible light, or of any other kind of electromagnetic radiation from radio waves to X-rays and gamma-rays. For example, assume, as in Figure 42, that light comes from a source, such as a lamp, or a sodium flame, at *A*, passes over a diffraction grating at *C*, and is recorded on a photographic plate *P*. If the wavelength is chosen suitably, we know we must expect to find dark and light fringes on the plate, as described in Chapter 3; these are sketched in the figure below *P*.

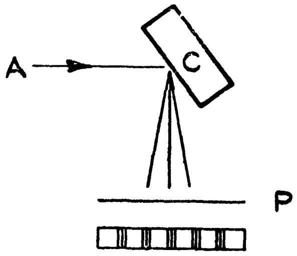


Fig. 42. Diffraction.

In Chapters 2 and 3 we have learnt how one obtains such diffraction patterns from the laws of electromagnetic waves. The photographic plate responds to the electric field, and the black lines, where the plate has been affected by the light, are places where the electric field is strong. The blackening of the plate (as well as the response of our eye) depends on the square of the electric field strength *E*, which, as we saw in Chapter 2, is also a measure of the electric energy density.

Now we know from the quantum theory that the light wave consists of light quanta, or photons, and the blackening of the plate is therefore the result of the action of a large number of photons. We can be sure that these photons do not influence each other in their travel. This follows because even near a moderately bright light source the number of quanta that are passing at any given time within a given space is not very large. (At a distance of one metre from a one-watt light source, there are only 1,000 photons in each cubic cm.) On the average their distances from each other must therefore be, on the atomic scale, very large. This is in accord with the observed fact that the diffraction pattern does not change if we reduce the brightness of the light, provided we expose the photographic plate correspondingly longer. What should we see if we made the light source so weak that we in fact record the passage of a single photon? This can give us only one black spot on the plate. If we repeat this often enough, all these black spots together will combine to make up the diffraction pattern predicted by wave theory. We may compare this situation with a marksman firing at a distant target. Each shot leaves a small mark on the target. These do not all appear in the same place because of errors in aiming, inaccuracies in the rifle, in the charge and in the effect of the wind on the bullet. For a single shot we cannot predict exactly where it will hit, but we know that many shots will build up a pattern which, if the marksman knows his job, will be densest near the centre, with only a few marks out on the edge of the target. We say then that each shot has a high probability of hitting near the centre.

In the same way we may say that, for a single photon, the diffraction pattern represents the probability of it making a mark in a particular place on the photographic plate. Since we had found that the blackening of the photographic plate depends on the square of the electric field, we conclude therefore that the square of the electric field in the light wave at each point determines the probability of a photon acting at that point on the photographic plate, and this conclusion helps to set up the relation between the wave and the particle description.

There is nothing new or surprising in the fact that, in studying the propagation of photons, we should get involved with probabilities. Just as in the case of the rifle shot it is not possible in practice to check the aim, or to investigate the various sources of error with sufficient accuracy to know the result precisely in advance, we could never hope in practice to follow the fate of a single photon with absolute precision.

But there is something surprising in this view of diffraction. In considering diffraction in Chapter 3 we had seen that, to obtain the wave scattered from the many lines of a diffraction grating, we had to add together the fields of the waves scattered by each line. If these acted in opposite directions, they were able to cancel each other, and this was essential in building up the interference fringes. In other words, if the electric field of the wave scattered from one line at some given place is E_1, and that scattered from another line is E_2, then the combined field is $E_1 + E_2$. The electric energy of the combined wave is therefore $(E_1 + E_2)^2$, which is $E_1^2 + E_2^2 + 2E_1E_2$, and therefore quite different from what we would get by adding the effects of the separate waves, which would give $E_1^2 + E_2^2$. At a place between two interference fringes, E_1 and E_2 may be quite large separately, but one of them positive and the other negative, so that they cancel out.

This possibility of cancellation by interference has no counterpart in the classical mechanics of particles. If the bullet from a rifle can reach the target in two ways, say either directly or by ricocheting off a rock, the number hitting any given part of the target can be obtained by finding out how many get there directly (even if the rock is not there) and adding the number that get there via the rock (even if the direct line is blocked). It would be absurd to expect that we should ever reduce the number hitting any part of the target by making both paths available. Yet this is what happens in the formation of an interference pattern.

What is new and unfamiliar about quantum theory is therefore that probability appears directly in the fundamental laws, and not merely as the result of our lack of knowledge of the mechanical details, and that the probabilities of separate alternative events do not always add in the simple way to which we are accustomed.

In the case of electrons, there is the same relation between the wave and the particle description. A single electron is not divisible, and if we look for it we always find it at some point, but we are not able to predict where it will turn up. All we can find out from the study of electron waves is the probability of finding the electron in any given place. Since the diffraction pattern which we observe in electron diffraction is quite similar to that found in the diffraction of light or X-rays, the probability must again be the square of a quantity which may be positive or negative, to give the possibility of interference. This quantity, which for electrons plays the same part as the electric or magnetic field for photons, is called the wave amplitude (or, using a more mathematical term, "wave function") and is usually denoted by the Greek letter ψ ("Psi"). This amplitude is what we plotted in our Figures 40 to 41. The probability of the electron turning up at some point is then equal to the square of the wave amplitude ψ at that point. For example in the case of the electron moving in the attractive field of force of Figure 40, it is most likely to be near the centre, if its energy is that of the lowest energy level, which belongs to Figure 41a. If it is in the second energy level, corresponding to 41b, it is unlikely to be just at the centre (where the wave amplitude goes through zero) but likely to be a little way on either side, etc.

Here again, the fact that probability enters into the fundamental laws and is not due to our lack of knowledge of the details, is illustrated by considering a simple case of interference, for example that shown in Figure 43.

A beam of electrons arriving from the left falls on a screen S which stops electrons, but in which there are two slits, s_1 and s_2. Behind the screen there is again a photographic plate, P. On this we again observe a diffraction pattern due to the interference of the waves which have passed through the two slits. Consider, in particular, a point p_1 on the plate at which there is a diffraction minimum, i.e. at which we find no electrons. If we were allowed to use the "classical" mechanics of Chapter 1, we should reason that, to get to the plate, the electron has to pass either through the slit s_1 or through s_2. If we block, say, s_1, this would stop electrons which were going to go through it, but would not affect those going through s_2. Suppose therefore we repeat our experiment, first with s_1 blocked and s_2 open, and then again with the same electron beam, and for the same length of time, with s_1 open and s_2 blocked. By the ordinary ideas of mechanics we ought then to get the same result as by doing it once with both slits open. But in fact we get a different answer. With either slit open, electrons will be able to reach the point p_1, and in each of the separate experiments the photographic plate is blackened at p_1. But with both slits open, no electrons are recorded at this point. This discussion makes it clear that we cannot explain the probability statements of quantum

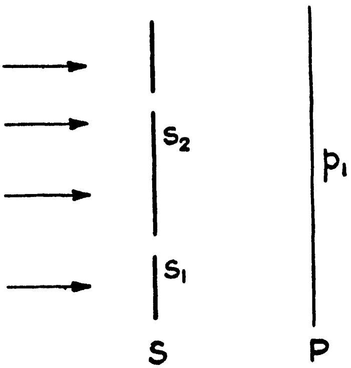


Fig 43. Interference from two slits.

theory by applying the laws of chance to any kind of ordinary mechanical motion. The statement that the combined probability of an electron getting to p_1 with both slits open is zero is one of the basic facts peculiar to the new situation.

At this point, the reader is likely to remark that, if closing one slit spoils our diffraction pattern, we might try to be less drastic. There might be other means of finding out whether the electron had gone through one or the other slit, without affecting its motion. If we had a sufficiently powerful microscope, we might just look for the passage of the electron through the screen and see whether it had passed through one slit or the other. It would be very surprising to find that the fate of all the electrons which had passed through slit s_1, was influenced by the presence or absence of the other slit.

But this is something which is easy on the scale of everyday life and difficult on the atomic scale, difficult not only for technical reasons but for reasons of principle. In talking about light, we have already noted that, because of the wave nature of light, we can "see" an object in the ordinary way only when we may regard the wavelength of the light as negligibly small; through a microscope of such magnification that it would make objects smaller than the light wavelength visible to the eye, they would appear as small blobs of the size of a wavelength because of diffraction effects. This means that, for our purpose, we would have to choose "light" of a wavelength less than the distance between the slits s_1 and s_2, in order to distinguish electrons passing near the one and near the other. But this distance must also be comparable with the electron wavelength in order to get the diffraction effects that we are talking about, and for electron velocities of practical interest this means "seeing" them by means of X-rays or gamma-rays. But in that region the quantum nature of the light is of importance. We can see an object only if some light has been deflected by it into our eye or into some other instrument, and therefore at least one light quantum must have been deflected by the electron. But then, the electron itself will have been deflected, i.e. its velocity changed because of the Compton effect. This will change its wavelength sufficiently to destroy the diffraction effect. This is only one possible method by which we might try to discover through which slit the electron is passing. One can discuss other possible methods but one needs quantum effects of a similar kind in all of them, and the answer always comes out to be the same: we cannot decide through which slit the electron passes, without disturbing its motion, and thereby destroying the diffraction pattern on the photographic plate.

The situation can be summed up by saying that, in order to obtain a diffraction pattern it is essential that we leave the electron undisturbed so that we shall be unable to decide through which of the two slits it has passed.

In the old mechanics it would have been reasonable to argue: if the electron passes through s_1 we can work out the possible orbits. If it passes through s_2 we get different possible motions. To obtain the complete result we must add the blackening of the photographic plate produced by electrons going one way to that of the others. In wave mechanics this reasoning fails because in this situation it is quite essential that we have *not* carried out any observation which would allow us to decide where the electron has passed through the screen. The question through which slit the electron has passed is therefore not one to which the answer can be given by any real physical observation.

It is not easy to get accustomed to arguments of this kind because they are so far removed from anything we experience when watching the motion of objects on a larger scale, objects which we can see or feel directly. When we watch a game of tennis we can see the tennis ball describing an arc through the air. When we see a rifle fired at a target, we cannot see the bullet, but are accustomed, nevertheless, to think of its motion following a similar arc (only much less curved) much more quickly; since a high-speed camera or other instrument can confirm this, we are in that case justified in extending our picture from the visible tennis ball to the invisible bullet. We have tended to generalize this to all objects, though this means applying the lesson of our experience to motion on a scale many million times smaller and to objects whose mass is 10^{27} times smaller than that of a bullet. But we ought not to be surprised if, so far outside our practical experience, we should in some place meet a new situation.

THE UNCERTAINTY PRINCIPLE

This problem raises for the first time the considerations expressed in the "uncertainty principle", which was formulated by Heisenberg and elaborated further by Bohr. This principle says that there is a limit to the accuracy with which we can study the motion of a small object. It is possible to find out where, at any particular moment, the object is located. Suppose we make any observation which locates the object within a distance *d*. In principle, we could make *d* as small as we like, i.e. locate the particle with as great an accuracy as we like, by using a "microscope" which operates with "light" waves of a sufficiently small wavelength. But in doing so we must, of necessity, expose the object to a collision with light quanta, which will change its momentum by an amount which may be as large as h/d, where h is Planck's constant. This disturbance is greater the smaller the distance d. Therefore, even if we had previously known the momentum of the particle accurately, it will now be unknown by an amount of h/d and (if relativity is unimportant) its velocity will be unknown by h/md, if m is its mass. Conversely, if we know to start with where the particle is located and make an observation which will tell us its velocity to high accuracy, this must, of necessity, involve such a disturbance of its motion that in the end its position will be uncertain by a corresponding amount.

The uncertainty principle says that it is generally impossible ever to know the position of an object to within a distance d and, at the same time, its momentum to better than a possible error h/d.

The uncertainty principle limits our possible knowledge of the state of motion of a particle to just such information as can be expressed in terms of a wave. Here the discussion of wave packets at the end of Chapter 3 is relevant. A wave packet like that sketched in Figure 24a has limited extent in space and this, remembering that the square of the wave amplitude gave us the probability, means that the particle is certain to be in a small region of extension l. At the same time, the wave packet is a mixture of waves of different wavelengths corresponding to different particle velocities so that we know the momentum of the particle to be between $\frac{h}{\lambda_1}$ and $\frac{h}{\lambda_2}$ where λ_1 and λ_2 are the longest and shortest wavelengths respectively. We saw in Chapter 3 that such a wave packet could be constructed only if $l\left(\frac{1}{\lambda_1} - \frac{1}{\lambda_2}\right)$ was at least about unity. Remembering that $\frac{h}{\lambda}$ is the momentum, this is at once seen to amount to the same as the uncertainty principle.

It follows that a sharply defined orbit of any particle, for example, of an electron, can never have a physical reality. If we try to observe the position of the electron with perfect accuracy, this would result in a complete uncertainty of its momentum, and therefore a complete uncertainty as to where it would be a little later, or even in what direction it was travelling. This must apply to any object, even to a tennis ball, but, of course, we would never attempt to locate a tennis ball with such extreme accuracy. Even if we knew its position to, say, one-millionth of a centimetre (which we could never achieve) and its velocity to one-millionth of a centimetre per second (which is also impracticable), the combined accuracy would still leave an enormous margin to spare by comparison with the uncertainty principle. Here we may therefore talk about an orbit without being troubled by any ambiguity. For as light a particle as an electron and within atomic dimensions, the position, however, is different. If we locate an electron to within one atom, i.e. to about 10^{-8} cm., then the corresponding uncertainty in its velocity is about 10^{8} cm./sec., which is just about the normal speed of an atomic electron. If we fix, even approximately, one point of an atomic orbit, we are disturbing the motion sufficiently to alter the orbit appreciably. One can, in fact, show that to distinguish by position and velocity two orbits which had the energies of two adjacent energy levels of the atom, would require an accuracy not permitted by the uncertainty principle. This is of interest because it removes the difficulty which we had found in trying to understand how, in the absorption or emission of light, the electron could change from one quantum state to another and how its motion during that intervening period was to be pictured. The answer we may now give to this question is that we may picture this motion in any way we like, but that such a picture would be idle speculation since there is no way of observing the details of the motion during such a period. Any observation of the orbit which had sufficient accuracy to distinguish one quantum level from the next, would have to cause a drastic interference with the process itself.

We see therefore that the contradictions that troubled us in trying to reconcile particle and wave aspects of both light and matter, arise always when we try to ask questions to which the answer cannot be supplied by any observation which does not itself violate the laws of quantum theory. The difficulties would remain if we insisted on asking such questions as, whether in the diffraction of electrons (Figure 43), the electrons had "really" passed through one or the other slit. The belief that such a question should necessarily have a reasonable answer derives from our experience of much larger objects, whose motion we can follow visually or in other ways without disturbing it. To make the laws of quantum theory consistent, it is important to accept the fact that such a question may have no objective answer, just as in relativity we learnt that there may be no objective answer to the question which of two distant events was the earlier.

One of the consequences of the uncertainty principle is that it changes our outlook on the law of causality. By the law of causality, one usually means the statement that the laws of physics determine the fate of a physical system completely, provided that all the relevant information about it is known at one particular instant of time. As we had learnt in Chapter 1, this is certainly true of any purely mechanical system. If, in the solar system, for example, we observe on one day the positions and velocities of all the planets in relation to the sun with perfect accuracy, we could calculate their complete motion at any subsequent or earlier date. In this instance, in fact, astronomers, while not achieving perfect accuracy, can work out the positions of the planets over very long periods indeed. The same is true of the laws of electricity. Assume we know at one instant the state of the electromagnetic field precisely and the motion of any charges contained in it; then the laws of Maxwell explained in Chapter 2 would permit us to predict its further fate. Quantum theory does not deny the statement that we could find the behaviour of our atoms or other mechanical systems at all times if we were given all the information for one particular instant, but it shows it to be quite meaningless because the condition which it assumes can never be met; we cannot ever know the state of the system to an accuracy better than that permitted by the uncertainty principle.

This is not to say that in quantum theory the laws of physics have become any less certain or less absolute. It is merely that the information to which the laws refer is of a different nature. In some cases they still allow some observations on individual particles to be predicted with certainty. This is always the case when we are dealing with a conservation law. For example, if a mechanical system is isolated and prevented from interacting with others, and from emitting radiation, its energy remains constant. If therefore we observe its energy at one time, and find it in one of its possible energy levels, and then make an energy measurement at a subsequent time, we can be certain to find it in the same level. But in general we are concerned only with probability, and thus an electron wave gives us the probability with which the electron is found in any part of space. However, this is still a precise statement, because we can reproduce the same electron state many times over, and if we look for the position of the electron every time and keep a record of it, the number of times the electron has been found in different parts will, after a long run of observations, correspond exactly with that predicted by the electron wave. We need, not one observation, but many to test the laws of quantum theory, and in that respect the laws are similar to those of the kinetic theory of heat, if for a different reason.

In principle, the laws of quantum theory apply to all objects however heavy, but, as already stressed, the corrections to the old mechanics of Newton are quite unimportant when we are dealing with large objects. This may be expressed by saying that we may have a wave packet which is small enough for its size to be negligible and yet will consist of waves of sufficiently similar wavelength to be moving with uniform velocity without spreading out. We have seen in Chapter 3 that, in the case of light waves, when we were concerned with dimensions much larger than the wavelength, we could think of light beams and light signals as travelling in definite paths and this approach to light had been called geometrical optics. We therefore see that the old mechanics of Newton bears the same relation to wave mechanics as geometrical optics to wave optics. Because of this analogy it was, in fact, at one time, believed that one could regard electron waves as physically real, so that an electron was just the same thing as a wave packet. However, a moment's reflection shows that this could not be so. In the conditions when geometric optics is applicable, wave packets will travel like small particles in definite orbits or paths. But the fact that they have a definite extent, however small, and that they contain a spread of different wavelengths, however little different, means that if we wait long enough, such a wave packet is bound to spread out and therefore, if a material particle, such as an electron, really consisted of such a closely limited wave packet, then after some time the waves would become spread out. In an experiment, we should then be able to cut off one half of an electron from the other, in complete contradiction with their actual behaviour. Hence we are forced to retain the view that the waves determine only the probability, and thus the spreading out of a wave packet does not mean that the electron has spread but merely that our knowledge of its whereabouts has become more uncertain. After a long time this is not surprising since there was necessarily some uncertainty in its starting position and velocity.

So far, we have discussed the uncertainty principle only in connection with possible measurements of the position and the velocity of a particle. These are not the only observations we can make. Others provide answers to other questions we may ask about the motion. Another important observation is the measurement of the energy of a particle moving in a known field of force, for example, an electron in a hydrogen atom. In Newton's mechanics, we can imagine that we have observed the position and the velocity of the particle and we can then calculate from the one the potential energy, from the other the kinetic energy and add the two. In quantum mechanics we know that this will never lead to an accurate result for the energy, since, of necessity, either our knowledge of the position or of the velocity, or both, must be inaccurate. However, we can also find the energy directly. For example, since we know the energy levels of the hydrogen atom and the frequency of the radiation emitted in the course of the transition from one to the other, we may consider an atom which has just emitted a photon corresponding to the energy difference between the energy level with $n = 2$ and that with $n = 1$, i.e. between the "first excited" and the "ground state". We may have observed the photon after allowing it to pass through a prism so that we know its wavelength and frequency. We then know that the atom has made the transition from its first excited state to its ground state and therefore if we arrange to have it protected from further disturbing influences it is now certainly in its ground state, of which the energy is known.

On such situations, the uncertainty principle also imposes a limitation which can be expressed by saying that ϵt must be greater than h, i.e. Planck's constant, where ϵ is the limit of error in our knowledge of the energy and t is a time whose meaning will be explained presently. It is sometimes said that, in this rule, t is the uncertainty in our knowledge of the instant of time to which the observation refers. In other words, the idea is that we know that, at some time, the particle had a certain energy, but we do not know when this was, and this uncertainty should be larger if the energy is specified more accurately. This interpretation is, however, not correct as can be seen at once from our example of the hydrogen atom in its ground state. Here we have found the energy precisely, i.e. the possible error, ϵ, is extremely small and yet, since the atom will remain in this state as long as there are no disturbing influences, we can choose any later time we like and say at this instant the energy is known.

There are, in fact, two correct ways of interpreting the rule. The first is this: if the particle, or the atom, changes its energy under external influences, or because it can emit radiation, and if therefore its energy remains constant only for a time of length t, then its energy is determined only with such a margin ϵ that ϵt is greater than h. In this form, the rule does apply to a hydrogen atom. The energy of the ground state is exactly determined because when the atom is left undisturbed, it will keep the same energy indefinitely. But, in any excited state, the atom has the possibility of emitting light and it will therefore stay in that state only for a limited period. We cannot predict the length of that period exactly because here again the laws of probability come in, but, on the average, the time is known. Therefore the energy of any excited state or of any energy level of the atom other than the lowest one is not precisely fixed but unknown with an error of about $\frac{h}{t}$, where t is the mean lifetime of the excited state, or the time after which, on the average, a light quantum will be emitted.

Since the energy of the light quantum must be the difference between the energies of the atom before and after the emission, this, too, must be uncertain so that the light observed in a spectrograph should not be mathematically sharp but should cover a small range of frequencies. That this is correct follows from a simple argument. If we have a large number of hydrogen atoms and put on a disturbance, e.g. pass through the hydrogen gas an intense electric discharge, which excites a large number of the atoms to the first excited state, the gas becomes luminous and emits light. However, this light lasts only for a short period because after a time, t, most of the atoms have returned to their ground state and have ceased radiating. The light which we observe, therefore, is a wave train or wave packet of limited duration, as in Figure 24, and we have seen in Chapter 3 that such a wave train is composed of waves of a range of wavelengths or frequencies and the spread of the frequencies is just $\frac{1}{t}$ where t is the duration. This gives just the right spread of energies for the associated photons. It is in fact true that spectral lines always have a width. In an ideal experiment, where other disturbing influences are avoided, one finds the "natural width" of just the magnitude corresponding with the above argument.

The rule that ϵt is no less than h may also be interpreted in another way. Consider an electron moving in a hydrogen atom in a very large orbit so that we may describe it by a wave packet, with an extension small compared to the radius of the orbit and yet a velocity spread small enough for the whole packet to move on an orbit without immediately dispersing. Then it is also true that if we know the energy to within a possible error ϵ the wave packet must be spread at least over a distance vt, where v is the velocity of the centre of the wave packet. In other words, the time when the electron would pass any given point on its orbit is then uncertain by t, and it is again true that ϵt must be at least h. Now, for a large orbit, the time of revolution, i.e. the time it would, according to classical mechanics, take the electron to return to its starting point, is also equal to the period of the light which in classical physics would be emitted from such an electron. Since, for very large orbits, quantum laws become identical with the conclusions of classical physics (this is the content of the "correspondence principle", which was of great historical interest in developing the quantum laws), it is still true that the frequency of the emitted light is $\frac{1}{t}$, or the energy of the light quanta $\frac{h}{t}$. Therefore the energy difference between adjacent energy levels, which is the same as the energy of the light quanta emitted in a transition between them is approximately $\frac{h}{t}$, where t is the period of revolution. It follows therefore from our rule that, the position of the electron on its orbit must be completely unknown if the energy of the electron is known sufficiently well to distinguish one energy level from the next. This again shows that we must not attempt to describe the transition of the electron from one orbit to the next in terms of the classical motion.

To sum up the results of this chapter: we have seen that both light quanta and electrons must be regarded as particles, but that we cannot follow the motion or the behaviour of these particles from point to point as we had been accustomed to in earlier chapters. Instead, their behaviour is described by waves from which we may deduce the probabilities of the results of observations on these particles. In this way, we learn that there is no contradiction between the corpuscular and the wave properties of light or matter, for both of which there is simple direct evidence; and at the same time, we see that the wave nature of electrons explains the stability of atoms with certain energies, the size of a normal atom and the spectral lines of light emitted by gases or partly absorbed by a passage through a gas. In subsequent chapters we shall see further how these new laws give a complete account of the behaviour of atoms and of systems of atoms in all conceivable circumstances.

The price which we have to pay for solving these problems is to accept the limitation of the concepts of position, velocity and motion, to all of which our everyday experience and several centuries of classical physics had accustomed us, and the recognition that these concepts can be used unambiguously only to the extent to which they can be linked with actual observations which do not, in principle, contradict the laws of quantum theory. In particular, we are forced to renounce the possibility of a complete causal description of mechanical processes, i.e. of ever knowing the state of an atom with such accuracy that its subsequent fate was completely predictable in the way in which an astronomer can predict the future position of a planet in the sky. This does not prevent us from predicting the behaviour of bodies containing large systems of atoms with certainty, since the laws of chance coupled with the quantum laws of probability allow such predictions with all the accuracy we shall ever require, just as was the case in the kinetic theory of Chapter 5.

8

The Behaviour of Atoms

HYDROGEN AND HELIUM

THE last chapter gave a sketch of the new quantum laws, to which we had been led by the many apparent contradictions met in applying our knowledge of the laws of nature to the interior of an atom. The main illustrations which we used were: the evidence for the existence of light quanta from the photo-electric effect, the Compton effect, and from the light sent out by a hot body, the evidence for the wave nature of electrons (and other particles) from electron diffraction, and the fact that atoms had a definite size and did not collapse.

It should not be forgotten, however, that these are merely examples chosen from a vast body of evidence which, at every step, indicated a failure of the old laws, and that the new laws were not finally accepted until one had seen that they could account, in a natural way, for almost all the known facts about atoms and could predict further regularities, which were confirmed by new observations.

The price we had to pay for gaining this new insight was to abandon the description of the atom in terms of simple pictures based on a comparison with the motion of bodies on a much larger scale. We had to recognize that such ideas as the position and motion of objects were prejudices to which we had been led by our experience in handling objects in practical life. We had to accept apparent contradictions such as that an electron was, in a sense, a particle and, in another sense, a wave, and we had to learn that these were not really contradictions, provided we did not insist on asking questions to which no possible observation or measurement could ever provide an answer one way or the other.

The last chapter outlined the main principles of the new laws but, for brevity, several refinements were left out. In the present chapter, we shall have to add these further points and also to describe some very general conclusions which can be drawn from the quantum laws. This will provide us with a complete set of laws which can now be used in earnest to study the behaviour of atoms, and this chapter will outline the results of such a study. It will then be seen that all the behaviour of atoms, and of matter which consists of atoms, fits into the picture, and that the quantum laws cover these problems as completely as the old mechanics covered the motion of planets, or the mechanism of a watch, and as Maxwell's laws of the electromagnetic field covered the behaviour of an electric generator, of a radio set, or the propagation of light waves.

To start with, we shall discuss some of the problems which arise when, instead of considering a hydrogen atom, which contains only one electron, we go on to larger atoms containing more electrons. To do this properly is not easy and requires some new ideas, which will be explained later, but for the moment we shall look at the problem as if the electrons did not exert any forces on each other, but each one was moving independently of the other in a field of force. This is similar to the way in which we can talk about the motion of the earth, or some other planet, around the sun as if only the sun and that planet were present, forgetting the existence of others. However, in the mechanics of the solar system, the reasons for doing this are very much stronger than in that of the atom because the gravitational force between two bodies depends on their mass. Since the sun is so much heavier than any of the planets, the attraction of the sun on the earth, for example, is very much stronger than, say, the attraction of Mars or Venus on the earth, even at similar distances. On the other hand, in the atom, the forces are electric and depend on the charge of the particles. Now the electric charge of all the electrons in an atom taken together is equal in magnitude (though of opposite sign) to the charge of the nucleus. Therefore the forces on one electron due to the others are not really weak. Nevertheless, it will be useful to start with the simplified problem because some important points will emerge, and we shall see later how the results can be improved.

Consider then the atom which follows hydrogen in the list of elements. This is helium, whose nucleus carries a positive charge twice as strong as that of the electron or proton. An electrically neutral helium atom contains two electrons. If we then disregard the fact that each electron may disturb the other we expect each of the two electrons to be represented by a wave of the kind discussed before. Each of those waves then belongs to some energy level labelled by the "quantum number", n; the state of lowest energy, in which the electrons are most strongly bound, is obtained by taking for each electron the lowest-lying energy level, i.e. the one with $n = 1$. The energy of this level was given in the previous chapter as $-\dfrac{4\pi^2 me^4}{h^2}$ for hydrogen, where m was the electron mass, e the electric charge of electron and proton and h, Planck's constant. The negative sign indicates that the energy is negative, i.e. that we must do work to remove the electron. The occurrence of e^4 in this formula comes from Coulomb's law, which gave the force as $\frac{e^2}{r^2}$. In helium, where the nucleus has a positive charge of $2e$, the force is now $\frac{2e^2}{r^2}$ and therefore the lowest energy level of an electron near a helium nucleus is now $-\frac{4\pi^2 m(4e^4)}{h^2}$. This is four times as large as for hydrogen. In other words, the energy required to remove one electron from a helium atom should be roughly four times as large as for hydrogen. Actually, these numbers are exaggerated because we have taken no account of the repulsion between the two electrons. But roughly, the answer is right; it is very much harder to strip an electron from a helium atom, or to "ionise" a helium atom. For example, if helium gas fills a tube of the type used in advertising signs, an electric current will pass through it only if the electric voltage between the ends is considerably higher than would be sufficient for a tube containing hydrogen. The mechanism of such a discharge is that some loose electrons are accelerated by the applied electric field and, in colliding with some of the atoms, may knock electrons off them. This leaves positive ions, which will travel towards the negative end of the tube, while the electrons which have been knocked off will travel towards the positive end. On acquiring sufficient speed, they may in turn ionise further atoms. The discharge can maintain itself only if, in the course of their travel, the electrons have the opportunity of acquiring enough energy to ionise more atoms in collisions, and the voltage at which discharge starts can therefore be used as a measure of the energy required for ionisation.

We can also look at the higher energy levels of the helium atom. We know, from the discussion of the last chapter, that light may raise the energy of the atom to one of the higher energy levels, provided the energy of the light quantum just equals the energy between the old and the new levels. Now in the simple picture we have made, all the energy levels of an electron in helium should be four times those of hydrogen. To raise a helium atom to a higher level or, as we say, to "excite" it, we should need light quanta of about four times larger energy. Since the energy of the light quantum is Planck's constant times its frequency, we need light of four times the frequency of the light which can excite hydrogen. Again the number four is an exaggeration because of the repulsion between the electrons, but it is true that the "spectral lines" of helium, i.e. the light frequencies which can excite the gas and are therefore strongly absorbed in their passage through it, lie very much further in the ultra-violet, i.e. towards higher frequencies, than those of hydrogen. So far, therefore, our picture is in good agreement with the facts.

OTHER LIGHT ATOMS. THE EXCLUSION PRINCIPLE

But this simple explanation holds no longer if we go further. Consider next the atom of lithium, which has a nucleus with a positive charge of $3e$, and three electrons in the neutral atom. From our previous reasoning, we should expect that the electrons in lithium should be about nine times as strongly bound as in hydrogen, that lithium vapour should be much harder to ionize than hydrogen or helium, and that the kind of light which can be absorbed by lithium should lie further still in the ultra-violet than for hydrogen or helium. These statements are not true. Lithium is easier to ionize than hydrogen, and therefore much easier than helium. A very low voltage is sufficient to start an electric discharge in a tube containing some lithium vapour, and if light which has passed through lithium vapour is resolved in a spectroscope, the absorption lines begin in the red, rather than only in the ultra-violet part of the spectrum. In fact, a flame in which we place some lithium salt shows a very characteristic red colour which is due to the emission of the same line. The mechanism here is that, in the collision of fast-moving atoms in the hot flame, some lithium atoms have got excited to their first energy level above the normal state, and they can get rid of this extra energy by sending out light quanta whose frequency is that of red light.

The lithium atom therefore does not show the behaviour which our simple picture would have led us to expect, and some extension of our laws is therefore required. This difficulty is connected only with the presence of the third electron. Once we have removed this, i.e. once we consider a lithium ion which contains only two in place of three electrons, it behaves entirely in accord with our expectation. In particular the energy needed to remove a further electron is indeed considerably higher than to strip one electron off a neutral helium atom.

It was noticed that the behaviour of the lithium atom would be given correctly by our model if we assumed that the third electron does not occupy the lowest energy level ($n = 1$) but the next higher one with $n = 2$. First of all, in the equations of Chapter 7, the energy of each level contained the square of the quantum number, n, in the denominator. In other words, the energy for $n = 2$ should be only a quarter of that for $n = 1$. Also, the radius of the sphere over which the electron wave spreads goes as n^2, so that the wave of the third electron would be spread over a much larger region than that of the first two. Therefore, we should think of the third electron as moving in an electric field produced by the nucleus (charge $+ 3e$) surrounded very closely by the first two electrons carrying between them a charge of $- 2e$. Its behaviour should therefore be similar to the case of an electron near a small centre with only one positive charge. Its energy should therefore roughly equal that of a hydrogen atom in the state with $n = 2$. In other words, the energy required to remove the last electron from a lithium atom should be only one-quarter of the ionization energy of hydrogen. It is actually only between two and three times less than that of hydrogen, but in view of the crude picture we are using at the moment, this discrepancy is not surprising.

It follows, therefore, that the normal state of the lithium atom is an energy level which we should have expected to exist, but which we would not have expected to be the lowest one. It looks as if the last electron is not capable of getting into the state with $n = 1$ if there are two electrons in it already.

We now go on to still larger atoms. The next one is beryllium, with four electrons, and the facts again support the idea that the fourth electron, like the third, can only go into the second level, i.e. $n = 2$. Adding successively further positive charges on the nucleus and further electrons to make the atom neutral, so that in turn we are considering the atoms of boron, carbon, nitrogen, oxygen, fluorine and neon, we find that they all behave as if they had only two electrons in the state with $n = 1$ and the rest with $n = 2$. As we follow this sequence of atoms, the attractive force from the nucleus becomes stronger and stronger, and as a result the wave representing the outer electrons become smaller and smaller in extent so that the size of the atom goes down, and with the more concentrated electron wave in the stronger field, the ionization energy rises progressively. In particular, the last in this sequence, the noble gas neon, is again very hard to ionize. Its ionization energy is almost as large as that of helium, and larger than that of hydrogen; its absorption lines begin only in the ultra-violet.

The next atom is sodium, with a nucleus containing eleven times the electric charge of the proton, and surrounded by eleven electrons, and here the situation that we found in lithium repeats itself. Sodium is easy to ionize, in fact easier than lithium, and the behaviour of the last electron suggests in all respects that it has now gone into the third level, that with $n = 3$. As we follow the series of elements, we find this situation repeating itself.

How can we account for this behaviour? It was at first believed that the reason why we cannot put more than two electrons into the lowest level and not more than eight in the next, etc., was due to the repulsive forces between the electrons, but a closer mathematical study disproved this idea. If one looks at the problem, for example, of three electrons in the field of a lithium nucleus in its exact form without simplification, the answer comes out quite definitely that the state of lowest energy would be that of all three electrons having $n = 1$, though the attractive energy in this state would be somewhat reduced by the repulsion of the electrons. Why then is this state not found in nature?

It was recognized by Pauli that this fact is a new piece of basic information of the greatest importance, which cannot come out as a consequence of the quantum laws explained so far. We must instead regard it as a new and independent rule which has to be added to these laws.

To find the precise nature of this rule, we remember that we were told in Chapter 7 that to the nth level of an electron in the hydrogen atom, there belong n^2 different electron waves. In other words, there is only one for $n = 1$, four for $n = 2$, nine for $n = 3$ and so on. Now we have seen that in building the atom, there seems to be room for two electrons on the lowest level ($n = 1$), eight electrons on the next ($n = 2$) and so on, and it looks therefore as if there was room for just twice as many electrons as there are different possible electron waves. No more than two electrons can be in the state described by the same wave. We may say that no more than two electrons can be in the same state of motion. In other words, any possible arrangement of electrons which, by itself, would be compatible with the laws of wave mechanics, but which contains more than two electrons in the same state of motion has to be excluded. This is essentially the content of the "exclusion principle" suggested by Pauli, which now forms an important part of atomic physics.

The reason why it is just two electrons which may move in the same way, is connected with a new discovery which we shall presently discuss in greater detail, namely the "spin" of an electron. According to this, we must picture the electron as spinning about its own axis while it moves in the atom, rather like the way in which the earth is spinning about its axis while it pursues its orbit around the sun. As we shall see later, in any given electron wave, or given state of motion of the electron, there are two possible ways in which it may spin. This suggests that, for example, the two electrons in the normal state of helium spin in opposite senses and the same will be true of the two electrons occupying jointly any other state in a bigger atom.

If one allows for this spin, the exclusion principle says therefore that electrons may never be in exactly the same state, if a state is specified by the motion of the electron and its spin.

As I have already stressed, this exclusion principle cannot be derived from the quantum laws, but is an independent fact for which we shall later meet a great deal of convincing evidence. When we deal with the theory of the positive electron, we shall see that the exclusion principle is vitally connected with the other laws in the sense that without it the other laws would lead to rather weird and unreasonable consequences.

ANGULAR MOMENTUM

To understand the facts about the spin of the electron, and also some other important characteristics of atoms, we must understand the part played in quantum theory by the angular momentum.

We had met angular momentum in Chapter 1 as a useful concept in mechanics. It measured motion around an axis, as opposed to motion towards it, or away from it. The angular momentum of an object was constant if the object moved under a central force, i.e. a force acting towards a fixed centre. This helped us to obtain very simply some information about its subsequent motion. If several objects exerted forces on each other this did not affect their total angular momentum.

To discuss angular momentum in quantum mechanics, consider first a very simple academic problem, that of an electron free to move only on a circle of radius r, like a bead threaded on a circular wire. If there is no friction we expect its kinetic energy to remain constant. If we introduce a wave, as in the last chapter, its wavelength will be constant. Since the path is closed, a wave can fit in only if the circumference of the circle contains a whole number, say m, of complete waves. In Figure 44 a thin line is drawn in such a way that its distance from the circle indicates the wave amplitude.

Using the de Broglie law, $\lambda = \frac{h}{p}$, where p is the momentum of the electron, we find therefore that $p = \frac{h}{\lambda} = m\frac{h}{2\pi r}$. This result is similar to what we found for the case of a box, in that only selected velocities, and hence only selected energies are possible in this motion. For $m = 0$, the electron is not moving at all; the wave amplitude, and therefore the probability of finding the electron, is the same all along the circle. For any other m the electron may be

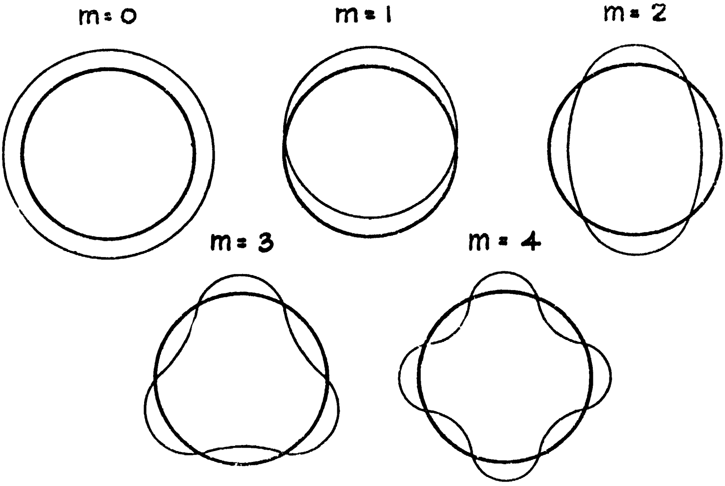


Fig. 44. Waves on a circle.

moving around in either direction, and to distinguish these cases we use positive and negative numbers for *m*.

Suppose the circle is horizontal, and take a vertical axis through its centre. Then the angular momentum about this axis is, according to Chapter 1, the distance from the axis, *r*, times the momentum of the particle in a direction at right angles to this distance. In our case, this is the whole momentum, so the angular momentum is pr, or $m\frac{h}{2\pi}$. The angular momentum comes out to be a multiple of $\frac{h}{2\pi}$, or a multiple of the "unit" of angular momentum.

The interesting point about this result is that the radius *r* has dropped out, and the result is the same whether the circle is large or small, though, of course, on a smaller circle the particles must move with a larger momentum, i.e. a larger velocity, to have the same angular momentum. This independence of the radius suggests that this result might be true even when the particle is not confined to move along a circle. And this indeed turns out to be true, though to prove it we would have to go into the mathematics of the problem. Consider first the case of a particle moving in a plane, for convenience of drawing, then it is still true that if we determine its angular momentum, we will always find it to be a multiple of $\frac{h}{2\pi}$

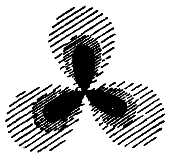


Fig. 45. Wave in a plane, $m = 3$.

and the wave describing the particle will then be somewhat as sketched in Figure 45. Here we have illustrated the case of an angular momentum of $3\frac{h}{2\pi}$ or, as we say for brevity, of three units. The shaded regions indicate where the wave amplitude is positive and in between it is negative. If we follow any closed line which surrounds the centre, the wave will change three times from positive to negative and back. At whatever distance we go around the centre, we thus always run through three complete waves.

This result, that the measurement of angular momentum always give exactly a whole number of units or an exact multiple of $\frac{h}{2\pi}$ is a result of the wave nature of the particle in just the same way as the fact established in the previous chapter that any measurement of energy will always show it to be exactly one of the energy levels of the atom in question. We shall see later, in connection with the experiment of Stern and Gerlach, that such a measurement of angular momentum is quite feasible in practice and is not merely an academic exercise.

Consider now the real case of an electron not confined to a plane, but moving in space in the neighbourhood of an attractive centre, as in the hydrogen atom. It still follows that a measurement of the angular momentum about any arbitrary axis through the centre can yield only a whole number of units. This, at first sight, seems to involve us in a contradiction. Namely, by following the definitions of angular momentum in Chapter 1, one can show that if we know the angular momentum of a particle about three axes at right angles to each other, e.g. about a vertical, an east-west and a north-south axis, we know it also about any other intermediate line. For example, if the angular momentum is large about a vertical axis and zero about two different horizontal axes, then it must also be zero about any other horizontal axis. For any line inclined to the vertical it will depend on its inclination, being zero when the line is horizontal and greatest when it is vertical. The situation described here is actually realized when the orbit of the particle is horizontal.

But this result, that on varying the direction of measurement the angular momentum should vary gradually between zero and the greatest amount, evidently cannot be reconciled with the conclusion that it always should equal a whole number of units. If it is not zero it can only be one unit, or more, but it cannot change gradually from zero to one.

This paradox is again resolved by the uncertainty principle. If we investigate possible methods of measuring angular momentum, we find that any measurement of the angular momentum about one axis must alter in an unpredictable way that about another axis, so that we can never know the angular momentum about several axes together. The connection of this result with the uncertainty principle in its simpler form is clear when we point out that knowledge of the angular momentum in all directions fixes the plane of the particle's orbit. This would be incompatible with the uncertainty principle since, assuming the plane to be horizontal, it would give us at once a knowledge of the height of the particle (namely, the same height as the attractive centre about which it revolves) and of its vertical velocity, namely zero, since the particle remains in the same horizontal plane. But we have seen before that knowledge of both the position and the velocity of the particle in the same direction are incompatible.

There is only one situation in which we can know the angular momentum in all directions and that is when it is zero. This would be the case in classical mechanics of an "orbit" which is simply a straight line through the attractive centre, and on which a particle moves back and forth, but the direction of this line would then be unknown; in any case this picture of the motion must not be taken too literally.

An important quantity in classical mechanics is the "resultant" angular momentum, M, which is defined by the statement that its square equals the sum of the squares of the angular momenta about three lines at right angles to each other, $M^2 = M_1{}^2 + M_2{}^2 + M_3{}^2$. It turns out that this resultant angular momentum does not depend on the direction of the three axes as long as they are at right angles to each other; if we choose them in such a manner that M, and M_2 are zero, then $M = M_3$. This is so when the third axis is at right angles to the plane of the orbit. In quantum mechanics, it turns out that the resultant angular momentum can still be defined in the same way and its square turns out to be, in the same units, always equal to $l(l + 1)$, where l may be 0, 1, 2, or any other positive number. The uncertainty principle does not prevent us from knowing both the resultant angular momentum (and hence l), and the angular momentum about one particular axis, which must be a whole number m of units. The angular momentum about any axis cannot exceed the resultant and therefore m cannot exceed l.

If we know the resultant angular momentum, and measure the angular momentum along a particular axis, we may find any of the results l, $l - 1$, $l - 2$, . . , $- l$ units, in all $2l + 1$ possibilities. For example, if $l = 1$, m may be 1, 0, or $- 1$.

We may think of the different states with the same resultant angular momentum as of orbits of the same shape but different orientation, like those shown in Figure 46. We note that, for $l = 1$ and $m = 1$, the angular momentum about the given axis is still less than the resultant which is the square root of $l(l + 1)$, i.e. $\sqrt{2}$ units, so we can never find the orbit to have its maximum angular momentum in the given direction. This is as it should be, since, if we did, we would know the plane of the orbit, which we had found to be impossible. That is why the orbits for $m = 1$ and $m = -1$ in Figure 46 have been drawn to revolve about an oblique axis. Whether this is inclined to the given line to the right or left or forward or backward or any other way must remain unspecified.

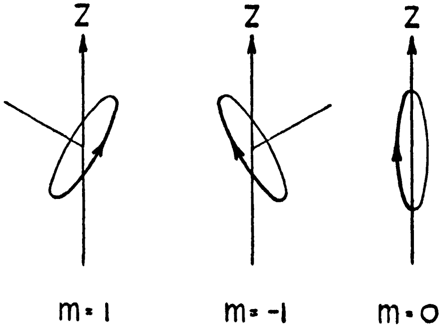
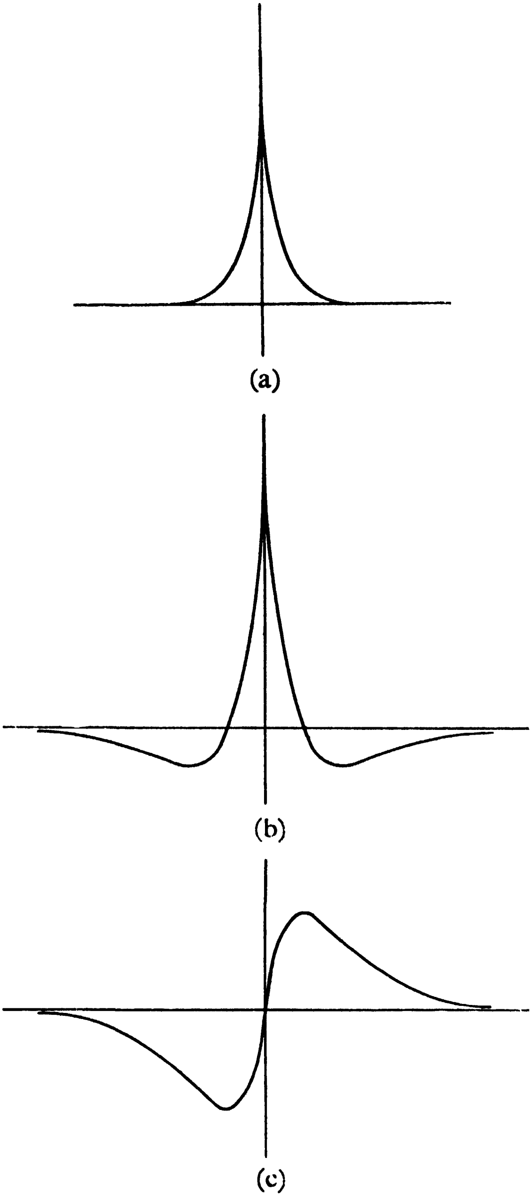


Fig. 46. How we might picture the orbits with $l = 1$ for motion around a centre.

We may now describe in some more detail the waves belonging to the lower states of an electron in a hydrogen atom. For $n = 1$, there is only one wave, and it belongs to $l = 0$ and consequently also $m = 0$. In this case there is no angular momentum. Consequently if we follow any line which encloses the centre, we do not follow any oscillation of the wave, the wave is the same in all directions. The wave, however, does depend on the distance from the centre and, in Figure 47a, we have plotted the wave amplitude along some line through the centre of the atom. We remember that the square of the wave is the probability of the position of the electron. In other words, the square of the height of the curve in Figure 47a tells us how likely we are to find the electron at some point along this line. In Figure 47b we have shown similarly a section through the wave for $n = 2$, still without angular momentum. Here the wave has a node, i.e. it is negative at large distances and positive close to the



(a)

(b)

(c)

Fig. 47. Waves in the hydrogen atom.

$$\left.\begin{array}{l}\text{(a). } n = 1, l = 0 \\ \text{(b). } n = 2, l = 0 \\ \text{(c). } n = 2, l = 1\end{array}\right\} \text{ radial sections.}$$

centre, and on the whole, it is spread out much farther from the centre, as corresponds to the larger mean radius, or to the fact that the electron is much less strongly bound. In addition to this state, there belong to the same energy three more states, all also with $n = 2$, but with $l = 1$. One of these is shown in Figure 47c. Now the wave amplitude is positive to the one side of the centre and negative on the other. A section along a plane through the centre for the same wave is indicated in Figure 47d where the lines indicate lines of constant wave amplitude as on a contour map. The line through the centre belongs to zero amplitude, to the right of this the amplitude is positive, to the left it is negative. Two other states for the same n and l are otherwise similar, but turned in different directions.

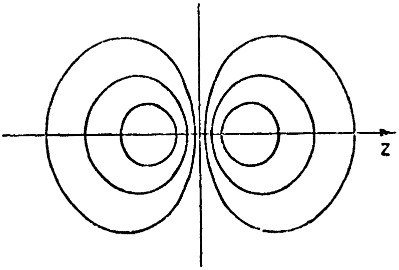


Fig. 47. (d). $n = 2$, $l = 1$—contour map.

THE ATOMIC BEAM EXPERIMENT. SPIN

The angular momentum of such an electron state can, in fact, be detected in practice because an electron revolving about a centre carries around an electric charge and behaves therefore like a small electric current in a little coil. As we saw in Chapter 2, magnetic fields exert forces on such a coil. We saw that a small coil of current produces the same magnetic field as a little magnet, a fact which is used in the "moving coil galvanometer", where the ordinary magnetic needle of the galvanometer, described in Chapter 2, is replaced by a little coil which is free to turn, and which carries an electric current. Now such a coil has a tendency to turn in such a direction that its magnetic field opposes that of the outer magnet. We conclude, therefore, that its energy is high when it is turned parallel to the external field and low when it is opposite. Therefore if the magnetic field is not uniform but is stronger in one place and weaker in another, a coil pointing the same way as the field will experience a force towards the weaker field, whereas one which is opposite to the magnetic field will be pulled towards where the field is stronger.

The same is true about atoms and this is exploited in the experiment of Stern and Gerlach which, in principle, is sketched in Figure 48. A tube is carefully evacuated so that practically no air remains in it, except for a small amount of gas which is allowed to escape from the chamber *C*, through very fine holes in the screens, *S*. The gas pressure is kept so low that each atom will travel in a straight line, since it has very little chance of colliding with another atom. Only those which happen to travel in the line joining the two holes will therefore emerge and they will travel down the centre of the tube along the broken line indicated in the figure. Such a narrow beam of atoms is known as an atomic beam. This beam is now passed between the poles of a large magnet, *M*, and these are shaped in such a manner that the magnetic field is much more intense near the upper pole, which ends in a knife edge, than near the broad lower pole. Hence, if any atom which has the electron rotating in the same sense as the current in the electro-magnet it will be pulled towards the weaker field, i.e. downwards, and if it revolves in the opposite sense, upwards. If the revolution of the electron is about a horizontal axis, then the atom will not be deflected at all.

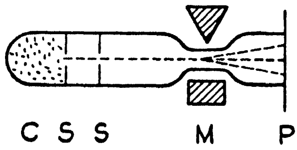


Fig. 48. Atomic beam experiment.

One can see that the expected deflection of the atom depends just on the angular momentum about the vertical line. Taking therefore an atom in which we have an electron with a resultant angular momentum corresponding to $l = 1$, then the angular momentum about the vertical axis may be 1, 0, or -1 units and we expect that some atoms will therefore be deflected upwards, some downwards and some not at all, resulting in the three broken lines shown in the figure. The atoms are intercepted on the plate *P* and the experiment is continued until enough atoms have accumulated there to make them visible, or to detect their presence by other devices. We therefore expect to find, in the case in question, three marks on this plate, one above, one below and one on the centre line. On the other hand, if we deal with an atom for which $l = 0$, i.e. with no angular momentum, this can carry no circular current and therefore will not be deflected at all. We then expect to see only one mark in the centre.

Now in some cases, the results are as expected. For example, helium, which has, as we saw before, two electrons in states with $n = 1$, and therefore $l = 0$, gives, as expected, no deflection. But hydrogen atoms, which equally should give no deflection, do in fact show two marks, and these two marks correspond to the deflection which would be expected for $m = 1$ and $m = -1$; but there is no centre line.

The explanation is that, as mentioned before, the electron seems to be capable of spinning about its own centre quite apart from its motion around the proton. We have seen that if the resultant angular momentum is $\sqrt{(l(l+1))}$ units then there are $2l+1$ possible states, so that for $l = 0, 1, 2 \ldots$ the atom has $1, 3, 5 \ldots$ possible states. Since from hydrogen we obtain two beams, the spin should have only two possible states, and this is obtained if we take l to be $\frac{1}{2}$. Then also the angular momentum of the spinning electron about any line should amount to either $\frac{1}{2}$ or $-\frac{1}{2}$. But as a magnet, the spinning of the electron must be twice as efficient as its revolution in its orbit because the experiment shows that, when the angular momentum about the field direction is $\frac{1}{2}$ the force is the same as that due to an orbit with $m = 1$.

Historically, the suggestion of the spinning electron was first put forward by Goudsmit and Uhlenbeck, not on the basis of the experiments described here but from a study of the Zeeman effect, which consists of the change of the spectral lines of a gas when a strong magnetic field is applied. The basis of the argument was very similar to the one outlined here, but its description would require more mathematics.

Thus the hydrogen atom has, after all, an angular momentum due to the spin of the electron, but helium does not, because the two electrons which move in the same wave must, by Pauli's exclusion principle, have opposite spins, and their angular momenta must therefore cancel.

Next we turn to lithium. The two inner electrons will still have opposite spins, and the angular momentum of the whole will now depend on which of the four possible orbits with $n = 2$ we select for the third electron. Now, while for hydrogen the four states with $n = 2$ have the same energy, this is no longer true for lithium or any other element. The reason can be seen by reference to Figure 47. If we compare Figures 47b and 47c, we see that in the first the wave is appreciable near the centre and therefore, in this state, the electron has a fair chance of getting close to the centre. The curve of Figure 47c crosses the horizontal axis at the centre, which means that the wave and therefore its square is zero at the centre. Here the electron tends to stay away from the centre. This is certain to be true in classical mechanics for a particle with angular momentum on account of the centrifugal force, as we saw in Chapter 1, and we now see it is still true in quantum theory. Now the attraction on the electron is much stronger near the centre compared with the hydrogen atom, because near the centre the screening due to the repulsive force of the two inner electrons is least effective. Therefore, if the two states have just the same energy in hydrogen, the first, with $l = 0$, will show stronger attraction and therefore have a lower energy when there are inner electrons present. Hence we expect the last electron of lithium to be in the state with $l = 0$ and, like in hydrogen, its angular momentum is only that of the electron spin. A beam of lithium atoms should in the atomic beam apparatus (Figure 48) again split into two, the deflecting force being just the same as in the case of hydrogen and this is exactly what happens.

The next atom, beryllium, has a second electron added in the $n = 2$, $l = 0$, state. To satisfy the exclusion principle, its spin must be opposite to that of the first. Hence the atom has zero angular momentum, and the beam should not be split. This again is confirmed by the experiment.

Now we come to boron. Since the levels with $n = 1$ and $n = 2$, $l = 0$ are occupied, the electron must go into the next level with $n = 2$, $l = 1$. The angular momentum of the atom will then depend on the direction of the spin of the electron relative to its orbit. Now which arrangement is most favourable is again a complicated question, which depends on the fact that both the revolution of the electron in its orbit and its spinning about its centre produce magnetic fields. If the interaction is worked out properly, it turns out that the state of lowest energy, and therefore the normal state of the atom, is obtained by the spin being as nearly parallel to the orbital angular momentum as, on account of the uncertainty principle, it can be. Therefore angular momenta add and since the orbit can have, at most, one unit about any given axis and the spin at most one half, the result is a maximum of $\frac{3}{2}$, and the normal boron atom therefore has $(2 \times \frac{3}{2} + 1) = 4$ possible states, the angular momentum about any axis being $\frac{3}{2}$, $\frac{1}{2}$, $-\frac{1}{2}$ or $-\frac{3}{2}$. A beam of boron atoms in the Stern-Gerlach apparatus should split in four; this again is true. The strength of the deflecting force can also be predicted. There is some complication of detail because the spin is twice as effective in its magnetic action as the orbit, but the answer can be worked out and is again confirmed by experiment.

Next comes the atom of carbon. Here we have two electrons in the $l = 1$ state. It again turns out they will arrange themselves in such a manner that both their angular momenta and their spins add; two orbits of $l = 1$ and two spins can add up to a maximum of three. A beam of carbon atoms would therefore split into seven (i.e. $2 \times 3 + 1$) branches. This has not, in fact, been done, since carbon vapour is not easily produced in practice, but there is no doubt that the experiment would again give the right answer.

We can go on in this manner until we come to fluorine, in which all but one of the places with $n = 2$ are filled. Instead of laboriously working out what each of the electrons does, it is easier to remark that, if we added one more electron, we should have filled all the places and then the angular momenta in any given direction would have to cancel, there being always two electrons with $m = 1$, two with $m = 0$ and two with $m = -1$, while three of the electrons must have spin of $\frac{1}{2}$ and the other three of $-\frac{1}{2}$ about the same axis. Since therefore the addition of one further electron would completely cancel the angular momentum, the angular momentum of the fluorine atom must be opposite and equal to that of a single electron in the $l = 1$ state. In other words, we get the same result as in the case of boron except that, for a given angular momentum, the direction of the equivalent electric current and therefore the direction of the magnetic field is opposite. This does not show up in the experiment of Stern and Gerlach but can be verified by the study of the spectra of such gases in a magnetic field.

In the noble gas neon, the last place with $n = 2$ is filled; the atom again has no angular momentum. In the following atom, sodium, the last electron has to go into the state with $n = 3$, $l = 0$, giving it characteristics very similar to those of lithium.

ELECTRON SHELLS, IONIC MOLECULES

We see therefore that, as we run through the periodic table of elements, the different energy levels get filled successively. It has become customary to use abbreviations for the states with the quantum numbers 1, 2, 3 . . . respectively, and they are usually referred to as the *K*, *L*, *M*, *N* . . . shells.

How these shells are gradually filled as we add more and more electrons is illustrated in Figure 49. This starts with the energy levels of hydrogen and shows one electron occupying the lowest one. The shaded region at the top indicates energies high enough to remove the electron from the atom; the distance from the lowest level to the beginning of the shaded region is therefore the energy required to ionize hydrogen. In the following diagrams for other atoms, it is important to remember that these are not drawn to scale. If they were, each diagram should be larger than the previous one, and, for example, in the case of sodium, the distance of the lowest level from the beginning of ionization should be a hundred times that for hydrogen and the distance of the last electron from the beginning of the shaded area should be almost as large as the whole picture for hydrogen.

The arrows on the symbol for the electrons indicate spin direction.

For any one electron it does not matter which way the spin points, but the diagram illustrates that electrons in the same state of motion must have, of necessity, opposite spins.

We have thus to picture the absorption of light by an atom in this way: If a light quantum arrives whose energy is just right to change one of the electrons to a higher energy level the atom can use up the energy of the light quantum, which then disappears. The atom is in an "excited state" and it will get rid of the excess energy either by emitting another light quantum, or, if there are other atoms close by, the energy may be shared out between several atoms and ultimately converted into heat.

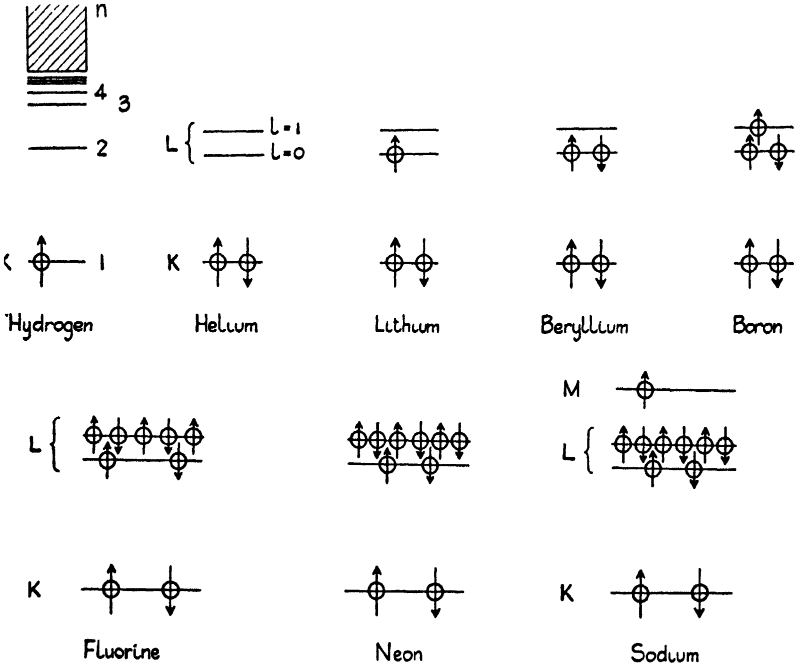


Fig. 49. Atomic shells.

The spectra of all atoms, i.e. the precise colours of the light which they are capable of emitting or absorbing, form a catalogue of great length, which has been compiled by spectroscopic work of high accuracy. It would go beyond the scope of this book to describe this in any detail, but it is important to know that it provides ample evidence in support of our general laws. It is not in practice possible to calculate from these laws the precise position of every spectral line of every element because when we are concerned with very many electrons the mathematical problems become prohibitive in complexity. But in all the cases which are simple enough to allow us to solve the equations, and in all those cases where one can obtain from the quantum laws some general rules about the nature of the spectra, the data are in agreement with the theory.

The same concepts can be used to interpret the behaviour of ions, e.g. of atoms which are either one electron short or have an extra electron. In particular, we know that the last electron in sodium is not very strongly bound because, with the K and L shell filled, it has to go into the M level with $n = 3$ and therefore has half the binding energy of the preceding electron. On the other hand, fluorine has one vacancy in the L shell and, although the atom is already electrically neutral, it is therefore possible to add another electron and this will be held with a considerable energy.

In fact, if an atom of sodium and an atom of fluorine get close together, the last electron of the sodium may be transferred to the fluorine atom, so that we are left with a positive ion of sodium and a negative ion of fluorine, which will attract each other. They will therefore stay together forming what the chemists call a molecule of sodium fluoride. In forming such a molecule from two neutral atoms, energy is released. It took energy to remove an electron from the sodium atom, but energy is recovered by putting it into the fluorine and more by the mutual attraction of the two ions. Evidently the same happens if, in place of sodium, we have any other alkali, i.e. any other atom containing one electron outside a closed shell. This is the case for any element following in the periodic table immediately after a noble gas. In the short list we have looked at in detail, lithium is another example. Potassium, which comes later in the periodic table following after the noble gas argon, is another. Equally, we may replace fluorine by any element which has just one electron missing to complete a shell and which therefore precedes a noble gas in the periodic table. Another example of this is chlorine, which precedes argon. Thus sodium chloride, more commonly known as common salt, is a typical example of this type of compound.

Because atoms of an element like sodium easily lose an electron to become positive ions, they are known as electro-positive, while atoms like fluorine are known as electro-negative. Hydrogen plays a double role, being both electro-positive and electro-negative. It may lose an electron, leaving just the hydrogen nucleus or proton, and then form compounds with an electro-negative element like chlorine. Such compounds have a strong tendency to replace the hydrogen by a more electro-positive element because it takes much more energy to strip the electron off hydrogen than, for example, off lithium or sodium. The combination of hydrogen and chlorine is, in fact, known as hydrochloric acid, and its corrosive nature is understood if we know that for example sodium chloride and neutral hydrogen have together a lower energy then hydrochloric acid and neutral sodium. On the other hand, hydrogen is also capable of accepting another electron, thus filling the *K* shell, and this leads to such compounds as lithium hydride, made of a negatively charged hydrogen and a positively charged lithium.

In this way the quantum laws for the atom also explain some of the processes in which atoms combine into molecules and in which elements can therefore form compounds. In this way some of the laws of chemistry are recognized to be consequences of the quantum laws.

OTHER TYPES OF MOLECULES

However, not all compounds can be explained in this way. Compounds in which each molecule contains one (or more) positive and one (or more) negative ion held together by electric forces are known as ionic compounds. Of a different kind of molecule formation, the simplest example is the hydrogen molecule, which contains two atoms of hydrogen. As was mentioned in an earlier chapter, natural hydrogen gas does not contain single atoms but atoms tied together in pairs to form molecules. These do not consist of ions, and in fact there is little reason to expect that of the two hydrogen atoms in a molecule one should be positively and the other negatively charged. How, then, does a hydrogen molecule hold together? To understand this, consider two hydrogen atoms approaching each other slowly as sketched in Figure 50. In Figure 50a, the atoms are still well separated; the dots indicate the nuclei or protons and the circles indicate the extension of the electron wave around each. In fact this electron wave does not have a sharp edge but diminishes gradually, but for the present purpose the picture is good enough. Since each of the atoms is electrically neutral it produces no electric field outside and there are no forces, either attractive or repulsive, between the atoms at this stage. In 50b the atoms begin to touch and in 50c they overlap.

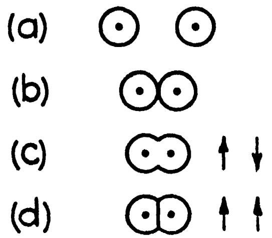


Fig. 50. Hydrogen molecule.

We had seen in the preceding chapter that the energy of the ground state of the hydrogen atom is a compromise between the negative (attractive) potential energy and the positive kinetic energy. If the electron wave were spread out more it would cover regions of weaker attraction, reducing the mean attractive potential; if it were smaller, there would be a stronger negative potential energy, but also the wave would have to contain shorter wavelengths, therefore higher velocities, therefore higher kinetic energy. Now if we have the position c, then the region in which the potential is strongly attractive, is larger and therefore each electron can spread out more and can have for the same attraction a longer wavelength and thus a smaller kinetic energy. Case c would therefore allow a lower energy.

However, we must remember the Pauli exclusion principle; and its effect is very different according to whether the electrons' spins are parallel or opposite. Consider first the case of opposite spins. In that event the Pauli principle does not prevent the electrons from moving in the same wave or in very similar waves and therefore they can both make use of the attractive region shown in Figure 50c. We conclude that, in that case, the energy of the electrons in case c is lower than that in a, in other words, that the energy of the two hydrogen atoms is much lower when they are close together than when they are far apart. Hence they attract each other.

But consider what happens when the electron spins are parallel. In that case, by Pauli's principle, the electrons must not move in the same wave, and in the attractive region of 50c only one electron can be in the wave occupying this region entirely. Putting this in different words, we may say that in this case we must divide the region into two parts, each of which belongs to one of the two electrons. This has been indicated in 50d, where the attractive region has been divided up. It is now evident that within the region where the potential is strongly attractive each electron has less space available than it has in a single atom. In this case, we expect therefore the energy to be higher than in separated atoms. We see in this way that two hydrogen atoms will attract each other if the electron spins are opposite, repel each other if the electron spins are parallel.

This explains why we get strong attraction only between two atoms at a time. We can form no hydrogen molecule containing three atoms, since there are only two possible directions of the spin. Indeed, suppose we start with a normal molecule of hydrogen, as in Figure 51, in which we have seen that the electron spins must point in opposite directions. If a further hydrogen

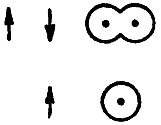
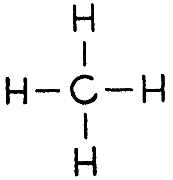


Fig. 51. Two hydrogen atoms are enough.

atom approaches, then the spin of this latter must be parallel to that of one or the other electron in the molecule, and therefore the space occupied by the two electrons in the molecule is largely barred to it by Pauli's principle.

These ideas fit in very well with the concept of chemical valency, which was introduced to describe the chemical behaviour of atoms. According to this, each atom has a certain number of "valencies" or possible bonds, and we can describe the possible compounds by joining up one such bond from each atom with that of another. The rule may be stated as if each atom had a number of strings hanging out from it, and a molecule were formed by tying up these strings to link the atoms. A stable molecule is then one in which no loose ends are left. Chemists write down formulae in which these valencies or bonds are indicated by lines, and a correct formula must have the correct number of lines starting from each symbol. Hydrogen has one such bond, and in this sense the symbol for a hydrogen molecule is H-H. According to the rule, this is a complete molecule without free bonds and should not be chemically active. This agrees both with the behaviour of hydrogen in practice and with our conclusions from quantum theory.

A more interesting case concerns the compounds of carbon (chemical symbol C), which has four such bonds, since the carbon atom has four electrons in the L shell. These may move in four different electron waves (characterised for example by $l = 0$; $l = 1, m = 1$; $l = 1, m = 0$; $l = 1, m = -1$; respectively) all their spins therefore being free to turn either way without violating Pauli's principle. Each of these electrons may therefore share part of its "living space" with that of another atom nearby, provided their spins are appropriately adjusted. Figure 52 shows the chemical symbol for methane, a gas found frequently in coal mines and oil wells and also contained in the usual commercial gas supply. Here four bonds have been drawn as starting from the carbon atom, and each links up with one from a hydrogen atom, so there are no loose ends left.



$$\begin{array}{c} H \\ | \\ H-C-H \\ | \\ H \end{array}$$

Fig. 52. The chemist's symbol for methane.

Compounds containing carbon and hydrogen in this way are of particular importance in the field of "organic" chemistry. The name indicates that compounds of this kind occur in great variety in the composition, as well as in the food consumption and waste output, of living organisms. It is not much of an exaggeration to say that the ionic compounds which we discussed first form the main basis of inorganic, and those connected with electrons sharing parts of the attractive field form the backbone of organic, chemistry.

The fact that noble gases like helium are chemically quite inactive is also immediately obvious. A helium atom does not easily lose an electron because this would have to be removed from the very low-lying *K* level, but it also does not like to accept an extra electron because this would have to go into the *L* shell where there would be practically no attractive force left, since the two *K* electrons which surround the nucleus closely will almost completely neutralize its field. Thus helium cannot form ionic compounds; but it also cannot share its attractive field with the electrons of any other atom, since there are in it already two electrons of opposite spin which utilize the attractive field already as well as the Pauli principle allows. In the case of two helium atoms, (or, for that matter, any two noble gas atoms), we always find repulsive forces when they get very close together because, by making the atoms overlap, we actually reduce the amount of space available for each electron and therefore force the electrons into states of shorter wavelengths and higher kinetic energy. The same applies also to the case of ions such as sodium and fluorine. A positive sodium ion and a negative fluorine ion have the electrons in the same states as two neon atoms. When they are so far apart that the electron waves do not overlap, the only force between them is the attractive force between a negative and a positive electric charge; but when they come so close together that the electron waves would overlap, this is outweighed by a repulsive force of the same kind as between two neon atoms.

SOLIDS, METALS, LIQUIDS

These considerations therefore lead naturally to an understanding of at least one typical form of solid matter. Consider a salt like the common salt, sodium chloride, which, as we have seen, consists of positive ions of sodium and negative ions of chlorine. From X-ray diffraction one finds that they must form a regular pattern of ions of which an example is shown in Figure 53. A basic part of this is a cube as outlined in the figure, with an ion at each corner, arranged in such a way that any two neighbours are of opposite sign. This cube is then repeated in a regular manner indefinitely. Now this is indeed a very favourable arrangement because in this manner each negative ion is closely surrounded by six positive ones and vice versa, so that ion finds itself in an attractive field. If we take into account only the electric attraction, the whole structure would immediately collapse because the attraction tends to pull each atom closer and closer to its neighbours. However, this will cease when the atoms begin to overlap, because then the repulsion of the closed shells comes into play. The normal state of solid sodium chloride will have the ions at such a distance that the electron waves are just beginning to overlap. In compressing it any further we would have to overcome the repulsive forces due to the overlap; in expanding it to a greater volume we would have to overcome the electric attraction between the ions. From our knowledge of the size of the electron wave in each atom it is therefore possible to find out what the distance between the ions, and hence the size of a given amount of sodium chloride, should be, and this answer again is found to be right. We can also compare the energy of the structure shown in Figure 53 with some other more complicated arrangement of the ions. Whereever this comparison has been carried out it shows that Figure 53 gives the lowest possible energy and should therefore be the form occurring in nature.

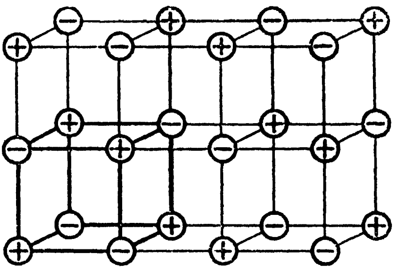


Fig. 53. Structure of sodium chloride (common salt).

This however is not the only type of solid matter. Another type owes its existence to the chemical bonds or valencies which we have discussed in connection with Figures 50-52. A typical example of this is diamond, which is pure carbon. Here the carbon atoms are arranged in a regular pattern, and each of them shares each of its four bonds with one of its neighbours. A solid like diamond poses a much harder problem to the mathematician than sodium chloride, but again we can go a long way towards understanding its behaviour.

Yet another interesting form of solid matter occurs in metals, such as sodium, aluminium, copper, silver, etc. These are all electropositive elements, i.e. their atoms lose one or more electrons fairly readily. Then the electrons are no longer attached to particular atoms but are free to move around throughout the whole piece of metal, no matter how large. The net result is again that the electrons can share the available space. When this argument is followed through quantitatively, the calculations show that this leads to an attraction which holds all the metal atoms firmly together in a solid piece.

In working this out one has of course again to allow for the exclusion principle. What this does can be explained by simplifying the problem very greatly. Let us ignore the forces which act on the electrons as they move through the metal and past the various atoms (now consisting only of the nuclei and their closed shells of electrons), so that we regard the electrons as moving freely in a large "box". Then each electron must be represented by a wave of such a wavelength that either a whole number of waves or a whole number and a half will just fit in the box. There is again a state of lowest energy for an electron for which the wavelength is twice the dimensions of the box and, according to the exclusion principle, only two electrons can be in that state. The next state is the one with the wavelength equal to the length of the box and there can be two electrons in this, and so on. To find the state of motion of all the electrons we must go on filling all these states until we have accommodated all electrons. It turns out that the fastest electrons that will be contained in the metal have a wavelength roughly equal to the distance between two adjacent atoms.

It was first pointed out by Fermi and by Dirac that even in such a large volume as a real piece of metal and even for a very large number of electrons, one has to allow in this manner for the exclusion principle, and the statistical study of large numbers of particles which are subject to the exclusion principle is now called Fermi-Dirac statistics.

At first sight, it is most surprising that such a principle should apply in this case, because we may apply our reasoning, say, to a piece of wire a few inches long. It is hard to believe that an electron at one end of this wire should be forbidden to have exactly the same velocity as an electron at the other end (unless they had opposite spins), because it is unreasonable that two electrons so far apart could influence each other in any way. This paradox resolves itself if we remember the uncertainty principle. Knowledge of the position of an electron, even to the extent of knowing in which half of the piece of metal it is, involves of necessity some uncertainty about its velocity. It can be shown that this uncertainty is just sufficient to prevent us knowing whether the electron is in one of the waves which fits into the given box or in the next one. Therefore no violation of the exclusion principle could ever be found in this case. In fact, instead of saying that the velocities of both electrons are exactly fixed and must be different, we can equally well specify the motion of the electrons by giving their velocity approximately and in addition some rough information about where each of them is. The accuracy of these statements must be so chosen that the uncertainty principle is satisfied. In that case we may say that if the electrons are in different places the exclusion principle does not restrict their velocities. Only if they may be in the same place (to the accuracy with which we have determined their positions) their velocities must be measurably different.

This fact, that the electrons in a metal are governed by the exclusion principle, or by Fermi-Dirac statistics, solved one of the major obstacles to an understanding of the behaviour of metals. It had been realized long ago from the high electric conductivity of metals that they must contain large numbers of electrons which can move about fairly freely. Now we had seen in Chapter 5 that if a space containing very large numbers of particles is heated, its energy should be increased by $\frac{3}{2}kT$ per particle, where k is Boltzmann's constant and T is the absolute temperature. If the number of electrons in a metal is comparable to the number of atoms, as is actually true, the electrons should therefore make a major contribution to the specific heat of the metal, whereas in fact the specific heat of a metal is very close to that of a non-metallic solid and can be accounted for solely by the motion of the atoms.

The same difficulty had arisen in Chapter 5 as regards the electrons inside any atom, but we have now learnt that such electrons are restricted to sharply defined energy levels, and at ordinary temperatures the distance between such levels is far greater than kT. The atom therefore just cannot accept the amount of energy which it should have according to kinetic theory if this is based on classical mechanics. This is not so in the metal because, since we are dealing with a very large box, the successive levels of an electron in such a box are extremely close together. However, the electrons fill all these energy levels completely up to a fairly high point. In Figure 54 the full lines indicate occupied and the broken lines vacant levels. Now if the metal is heated, each electron is quite likely to acquire an extra energy of about kT, but only a few electrons can take advantage of this. An electron near the top of the filled levels, say in the state marked A, may go to a state B if their distance

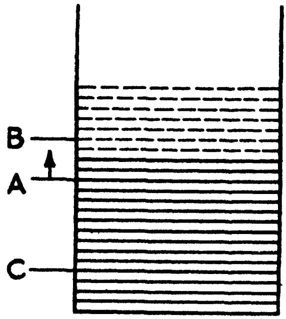


Fig. 54. Many electrons in a box (Fermi statistics).

is about kT. But, for example, the electron in the state marked C cannot be raised by the same amount because that would bring it to a level which is already occupied; and if this electron were to gain energy at all, a far greater increase of energy would be necessary. It follows from this argument that only a very small fraction of the electrons, those within an energy kT from the top of the occupied levels, can contribute to the specific heat and their number is in general too small to matter. In particularly favourable circumstances, however, this electronic specific heat can be observed.

This is only one example taken from the quantum theory of metals, which has developed into a wide and interesting field in which quantum theory has been used in many ways to bring order into an apparently disconnected list of observed facts. This field, however, contains also one of the major unsolved problems in the phenomenon of super-conductivity. Some metals are found to lose their electric resistance suddenly and spectacularly when cooled below a certain temperature, which is usually only a few degrees from the absolute zero. Once below this temperature their behaviour is also unusual in other ways. This is one of the few phenomena in the behaviour of matter for which no entirely satisfactory explanation has been found, though in the last few years suggestions have been made that seem to contain at least an essential part of the truth. However, it is most likely that we are not dealing here with a breakdown of the laws of quantum theory, but rather that this problem, which involves of necessity the presence of very large numbers of atoms and electrons, and depends very sensitively on their mutual interaction, is so complicated that we have not, as yet, been ingenious enough to find the right way of looking at it.

Another interesting problem about solids is their specific heat at low temperatures. We have seen in Chapter 5 that the specific heat of most solids at room temperature and above is $3k$ per atom, as required by the law of Dulong-Petit, which followed from kinetic theory. But as they are cooled to lower temperatures the specific heat falls and approaches zero as the temperature tends to the absolute zero. Figure 55 shows the characteristic behaviour of the specific heat of a solid. This can now also be understood.

According to the story outlined in Chapter 5, this specific heat represents the fact that the atoms can vibrate about their equilibrium positions. Each atom belongs to a certain position in the regular pattern, and the force of the neighbouring atoms tend to keep it there, so that when disturbed it will oscillate about that position. Now if quantum mechanics is applied to this problem we have a situation as in Figure 40, and we expect again to find sharp and distinct energy levels. Whereas kinetic theory would require the atom to vibrate on the average with an energy of kT, the least energy it can actually accept is that which would bring it to the first excited energy level. If this is much more than kT, i.e. at low temperatures, the chance of it picking up that much energy becomes very small.

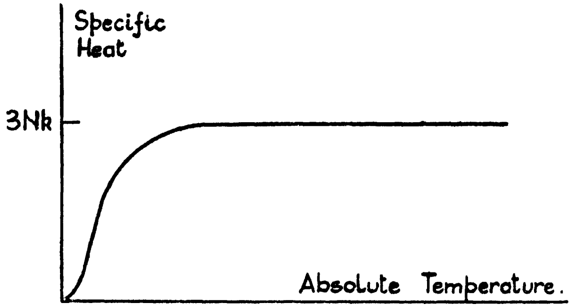


Fig. 55. Specific heat of a solid.

The situation here is very similar to that of black-body radiation, which was discussed in greater detail; for the reason given there, the heat energy is very much less than classical mechanics predicts. This simple explanation of the specific heat of solids is due to Einstein. The picture as presented here is over-simplified because I have argued as if each atom were moving independently of its neighbours. In a precise treatment one should take the motion of all its neighbours into account as well. The more exact description, which was first worked out by Debye, accounts very well for the known facts.

We have talked about gases and about solids but not yet about the liquid state. The general explanation of melting was already given in Chapter 5, and quantum theory does not alter greatly the picture there described. The mathematical treatment of liquids is more difficult than that of either gases or solids. The problem of a gas is simplified by the fact that molecules move independently of each other most of the time and collide only occasionally. That of a solid is simplified by the fact that each atom does not move far from its appropriate position in the regular pattern or "lattice". Neither of these statements apply to the liquid and we are left with the task of dealing with the complicated motion of very many atoms jostling each other continuously. The description of the liquid state in detail is therefore less complete than that of gases or solids.

One very exceptional case of a liquid is that of liquid helium. In our description of liquids in Chapter 5 we had attributed the melting, i.e. the breakdown of the regularly ordered pattern of the solid to the heat motion of the atom, but helium remains a liquid right down to the absolute zero and never solidifies unless a high pressure is applied. We know the reason for this. Helium atoms are very light, so that for moderate velocity their wavelength is very long; the attractive forces between them are extremely weak because helium is a noble gas in which there exist none of the types of chemical attraction which we have discussed before, and this is the reason why helium remains a gas down to extremely low temperatures and condenses only at a temperature of about 5 degrees from the absolute zero. This means that a curve like that of Figure 40a, to represent the potential energy of one helium atom, has to be very shallow indeed, and the small mass means that a greater kinetic energy is required to squeeze the wave into an attractive region of given size. It is therefore not surprising that the attraction may be insufficient to give any bound state at all, and that even for the lowest possible energy the helium atoms do not settle down into any regular pattern. In other words, the formation of a regular pattern is here prevented not by the heat motion but by the zero-point energy. Liquid helium is therefore different in nature from all other liquids, and it makes a fascinating object of study both for experiment and theory, but again the problem is so complicated that a complete analysis of its behaviour is still lacking.

These examples will have to suffice as illustrations of the way in which quantum mechanics has removed the difficulties and ambiguities found previously in trying to understand the structure of matter. For the behaviour of atoms in almost all situations met with in everyday life or in the laboratory the laws we have described up to this point have proved applicable and gave results in good agreement with observation. At the same time, the practical physicist can use them for developing or improving devices in which the properties of matter or of atoms are used for practical purposes or for further measurement.

PASSAGE THROUGH A POTENTIAL BARRIER

To conclude this chapter we shall discuss one more application, not for its own sake, but because it will help us to understand one result of quantum mechanics which will play an important part in later chapters.

To make the point let us return, for a moment, to the case of a particle moving under an attractive force as in the problem illustrated in Figure 40a, where, for a typical case, the electron wave was shown in Figure 41c. We have already noted in this connection that for motion with energy *E*, in Newton's mechanics, the particle would have to be confined between the points *A* and *B* since, outside of these points, the potential would exceed its total energy so that the kinetic energy would have to be negative. Now it is true that to the right of *B* and to the left of *A* the wave amplitude diminishes very rapidly and, bearing in mind the meaning of the wave as probability, it follows that the particle has only a very small chance of being found much beyond *A* or *B*. However, the point is that there exists a small chance of it getting there, i.e. of penetrating into a region for which, in the old mechanics, its energy would be insufficient. Does this mean that the law of conservation of energy is not valid in wave mechanics, or does it mean that the kinetic energy of a particle can, in fact, be negative?

The answer to this paradox is again supplied by the uncertainty principle. We cannot apply the law of conservation of energy to any actual situation by measuring the kinetic and potential energy accurately and adding them because for the one we must know the position and for the other the velocity of the particle, and the uncertainty principle says we cannot know both these things exactly. If we tried to catch the particle in the classically "forbidden" region to the right of *B*, we would have to determine its position fairly accurately (since it does not get far beyond *B*), and this would require a method of measurement capable of altering the energy of the particle sufficiently so that, when we have found the particle, we can no longer be sure that its energy is, in fact, less than the potential energy at the point where it has been found.

In the conditions illustrated in Figure 40a, the question is therefore somewhat academic, but it takes a much more practical form when the potential exceeds the actual energy of the particle only over a short distance and then decreases again. This leads us to the situation pictured in Figure 56, which illustrates a case in which a certain region repels the particle. We have here a barrier through which, in classical mechanics, the particle would not be able to pass were its energy less than the height of the barrier, for example, if its energy were given by the broken horizontal line in the figure. According to quantum mechanics, however, there is a small probability of the particle penetrating into the barrier and reaching the other side. Having once reached the other side, it may pass on. Figure 57 shows a wave arriving from the left on to the barrier shown in Figure 56. As in previous cases, there is again a point *B* where the potential and total energy are equal and where the curvature of the wave is reversed. The wave thus diminishes rapidly as it did to the right of *B* in Figure 41c, but this continues only until the point *A* is reached where the potential is again less than the total energy, and from there on an oscillatory wave is again possible though this is very much weaker than the one on the left. This picture therefore shows that the particle has a small chance of passing through the potential barrier.

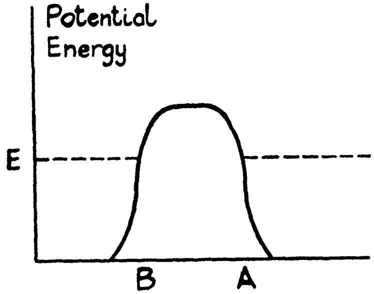


Fig. 56. Potential barrier.

The chance of this happening diminishes very rapidly with the height and width of the barrier. It is appreciable only when the width of the barrier, i.e. the distance from *B* to *A*, is not much more than the wavelength of a particle whose kinetic energy is just equal to the deficiency, i.e. to the height of the barrier above the broken line in Figure 56. If the width of the barrier, i.e. the distance from *B* to *A* is just equal to this wavelength the odds against the particle passing through the barrier are about half a million to one and the chance decreases rapidly with increasing thickness and height. It is therefore clear that such a passage through a potential barrier need be taken at all seriously only in the atomic domain. For larger objects, when we are concerned with shorter wavelengths and greater distances, the probability becomes quite fantastically small.

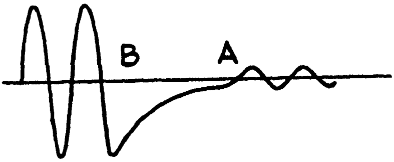


Fig. 57. Wave in the potential barrier of Fig. 56.

One typical case in which this process manifests itself is the so-called cold emission of electrons from metals. We have mentioned that an ordinary metal contains electrons which are responsible for its ability to conduct electricity. In normal circumstances, these electrons are prevented by an attractive force from leaving the metal, and work would have to be done against this force to remove an electron. Electrons may escape if the metal is heated to a high temperature, because by their heat motion some electrons may then have enough energy to overcome the attraction. This is utilised in the hot filament of an ordinary radio valve, which is heated just in order to make the escape of electrons possible. But alternatively, electrons may escape if the metal is placed in a strong electric field, for example by making it one of the two plates of a condenser as in Figure 11. The potential energy of an electron will then appear

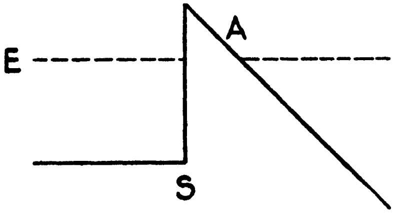


Fig. 58. Cold emission of electrons.

as in Figure 58, which shows a section across the surface of the metal. To the left of the surface *S* the potential energy is constant, and an electron with an energy marked by the broken line is free to move about. When it approaches the surface *S*, however, the potential energy rises rapidly because there is an attractive force which prevents its escape. Outside the metal, the potential diminishes again because of the effect of the attractive electric field and beyond the point *A* the motion of an electron of energy *E* is again possible. Here again there is a small chance of the electron getting across the potential barrier from *S* to *A*. This chance is greater if a stronger field is applied because the line representing the external potential slopes more steeply and the point *A* moves closer to *S*. This reduces the width of the barrier. The chance is also greater for a metal for which the barrier is less high, i.e. for which less energy is required to remove an electron.

This "cold emission" has been studied well and the variation of the rate at which electrons escape with the field strength agrees with the predictions of the theory.

Another useful way of looking at this situation is to express it in terms of the uncertainty relation between energy and time, to which reference was made at the end of Chapter 7. It was there pointed out that energy could not be determined accurately in a short time. If, therefore, the energy of the particle, instead of being constant, changed by a small amount for a short time only, we could never detect this. To put it crudely, a particle may "borrow" energy, and if the amount ϵ which it had borrowed, times the time, t, after which the loan is returned, comes to less than h, no measurement can show up directly the lack of balance. In this sense, we may interpret the passage of an electron through a potential barrier by thinking that the electron had borrowed the necessary amount of energy to get over the barrier; if the barrier is narrow so that the passage does not last long, it can return the loan rapidly enough so that nobody need be the wiser for it.

This possibility of "borrowing" energy will prove important in later discussion, and it is an argument which runs right through all problems in quantum mechanics. In discussing possible changes in the state of an electron or other particles we must allow for the possibility of such a change even where it appears to violate the law of conservation of energy, provided this state does not last long and the particles can, after a short time, return to another state in which the energy can be balanced.

One usually speaks of a state which can exist only for such a short time as a "virtual" state. Such virtual states will play a part later on.

9

Electrons at High Speed

NEGATIVE ENERGIES. A DIFFICULTY

IN developing the laws of quantum mechanics in Chapter 7, we had used the relation that momentum is mass times velocity. This is correct in the old Newtonian mechanics, but not for high velocities when the refinements of relativity must be taken into account. Now for an electron moving in a light atom, for example in hydrogen, this is not important since its velocity is much less than that of light. We saw that the binding energy of the electron in a normal hydrogen atom is $\frac{2\pi^2 me^4}{h^2}$ and one can show that the kinetic energy of the electron is equal to this. The kinetic energy also equals $\frac{1}{2}mv^2$, and therefore the mean speed of the electron is $v = \frac{2\pi e^2}{h}$. Comparing the electron speed v with that of light, c, we have $\frac{v}{c} = \frac{2\pi e^2}{hc}$.

This quantity is a pure number, i.e. it does not depend on whether we measure the speed in centimetres per second or in miles per hour, since it represents a ratio between two speeds. If we put in the known electron charge as well as the light velocity and Planck's constant, the number comes out to be about $\frac{1}{137}$. We see that the speed of the electron in a hydrogen atom is therefore less than one-hundredth of the velocity of light and the corrections due to relativity are therefore quite small.

There are, however, problems of a different kind. One is the movement of the electron in the K shell of a heavy atom, e.g. of lead, in which the charge on the nucleus is 82 times the charge of a proton. The binding energy of a K electron in lead is therefore $(82)^2$ times that of the hydrogen atom. Accordingly the same estimate of v/c would be 82/137. Hence the electron moves here with more than half of the light velocity and relativity is important.

It is also possible to accelerate electrons in the laboratory to similar and higher velocities. From the conclusions of Chapter 6, it is easy to show that, if an electron is to move with a velocity close to that of light, its kinetic energy must be greater than mc^2, where m is its rest mass. If the electron is to be accelerated to such an energy by an electric field, the difference in electric potential over the path of the electron must be $\frac{mc^2}{e}$ which comes out to be about half a million volts. It is customary to express the energy of an electron by stating the electric potential difference which would accelerate it to this energy, and we say that the energy of an electron may be so many "electron volts" or so many million electron volts (MeV). An electron accelerated to several MeV will therefore move with a velocity close to that of light.

This can be achieved either by passing an electric discharge through an empty tube containing a few electrons and applying an actual voltage of several million volts, or by some more ingenious devices which will be mentioned in the next chapter; electrons of similar energy are also found amongst the "beta" rays of radium and other radio-active substances. If we want to describe the waves belonging to such electrons, we must use the correct relativistic law relating energy and momentum which, in Chapter 6, we had found to be $\frac{E^2}{c^2} = p^2 + m^2c^2$, where E was the energy inclusive of rest energy and p the momentum. This relation is shown in Figure 59.

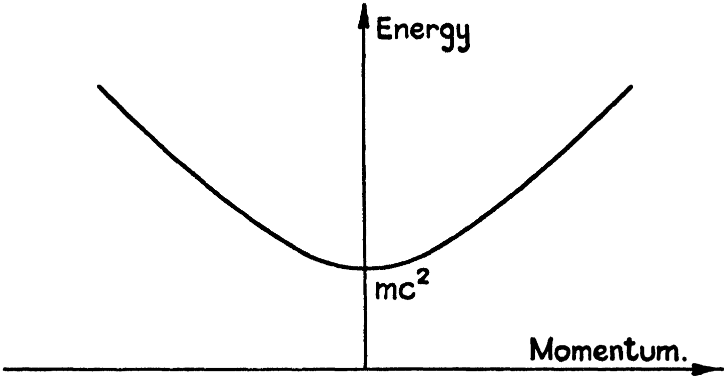


Fig. 59. Energy and momentum.

Here we meet a peculiar difficulty. The square of a negative number is positive. Since our relation only specified E^2, we might choose E either positive or negative. In other words, the two curves of Figure 60 are equally compatible with our relation between energy and momentum.

Could the lower branch also have any relation to reality? It describes a situation in which the total energy of a free particle is negative and decreases further as the momentum of the particle is increased, i.e. as it is accelerated. Such a particle would behave in a very queer way indeed. It would behave as if it had a negative mass; its acceleration would be in a direction opposite to that of the applied force. If we tried to pull it, it would run away. If such an electron was close to a normal electron it would repel it and therefore the normal electron would move away, but the repulsive force of the normal electron would make this hypothetical particle move towards it. The two would therefore keep accelerating in the same direction, the increasing kinetic energy of the normal electron being balanced by the increasingly negative energy of the other.

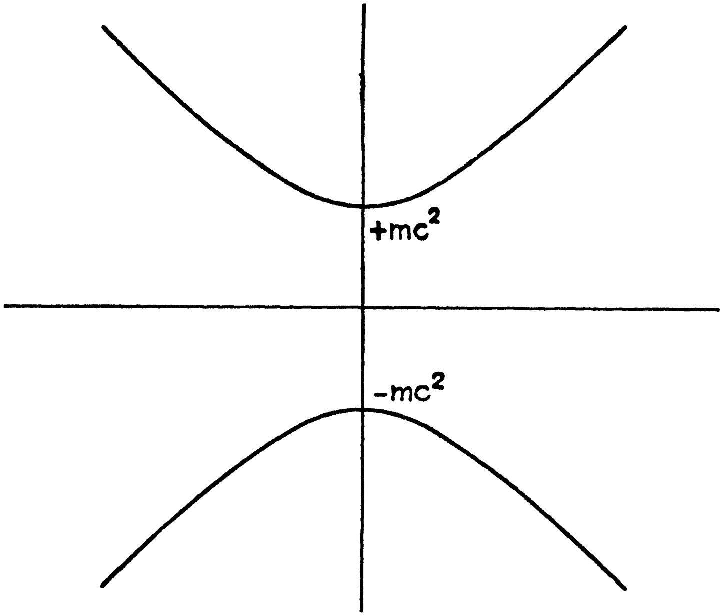


Fig. 60. Energy and momentum; what does the lower branch mean?

This behaviour is unlike anything we have ever met and if, in fact, such particles existed, matter could never settle down to any steady state, since these particles could always create major disturbances by starting to move fast and thereby liberating energy.

Now as long as we are not concerned with quantum mechanics, this difficulty is not serious. Under the influence of any reasonable force, the energy of a particle changes only gradually; it will therefore be impossible for an electron to jump from the upper curve of Figure 60 to the lower curve. Provided we can be sure that we start off with only normal electrons, the existence of the lower curve is therefore of no importance.

But in quantum mechanics things are different. We can no longer argue that energy must change gradually, since energy cannot be measured exactly in a short time. In fact, in an atom, which is capable only of selected energy levels, we know that it can go from one level to another with the emission or absorption of light and we are unable to account for the details of what it did in between. For the same reason if the laws of quantum mechanics allowed, in principle, waves which corresponded to the lower curve of Figure 60, we should have to be prepared for the electron to jump from the upper to the lower curve, just as it can jump from an excited to the ground state of the hydrogen atom. The difference in energy will, in general, come out in the form of light.

Actually, a jump from the upper to the lower curve of Figure 60 could not be accomplished with the emission of a single photon, since the energy to be disposed of is always greater than c times the change of momentum, whereas the energy of a photon is just c times its momentum. However, it is possible to balance energy and momentum for the emission of two photons. What is more, from any given state on the upper curve the electron can get to any arbitrary point on the lower curve and this extends to infinitely high momenta either way. If we attempt to work out the probability per unit time of this happening in any given time, we find an infinitely large answer. This would mean that the electron could stay only for an infinitely short time on the upper curve and would at once drop to the lower one. We should therefore never see any ordinary electrons but only the queer ones of negative mass.

SPIN. POSITRONS

Dirac attempted to solve this difficulty. He set out to find a law for electron waves which would be different from the generalization of Schrödinger's equation to the relativistic case, in the hope of finding a law which would allow only the upper curve of Figure 60 and not the lower. At the same time, he also tried to remove some mathematical difficulties which had been troublesome in the interpretation of the relativistic wave equation.

In this way, he was led to consider the description of the electron in terms not of one wave, but four. This is analagous to the case of electromagnetic waves, which, as we saw in Chapter 2, are described by the electric and magnetic field strengths, each of which has direction and magnitude, and can therefore be specified by its projections on three different axes, so that an electromagnetic wave is fully specified only by six numbers.

The consequences of this new fourfold wave equation were most interesting. It turned out that they gave the electron besides the angular momentum due to its motion, a further angular momentum, which could be either one-half or minus one-half unit about any axis. Thus the spin of the electron proved to fit naturally into this relativistic picture. In addition it also turned out that the magnetic action of this spin was twice as effective as that of the motion, so that the magnetic action of a spin with half a unit of angular momentum is as large as that due to an electron revolving in an orbit with one unit. This we had already noted as an experimental fact in Chapter 8.

It is by now evident that this fourfold wave equation gives a much more natural description of the electron, and further investigation has shown that the scattering of X-rays by fast electrons (i.e. the Compton effect of Chapter 7), the ejection of electrons from atoms by X-rays or gamma-rays (i.e. the photo-electric effect at high energy), the deflection of fast electrons by atoms, and many other phenomena came out from Dirac's equation in precise agreement with experiment.

It looked, at one time, as if the Dirac equation was the only consistent way of describing a charged particle. If so, we could have predicted that the electron must have a spin even if we had never noticed it. We now know that this view is not correct. It is possible to deal with the wave equation for spinless particles and, as we shall see in Chapter 11, spinless charged particles do indeed exist.

However, the difficulty about states of negative energy still persisted. Like the simple wave equation, Dirac's equation also admits energies corresponding both to the upper and the lower curve of Figure 60. As long as this difficulty remained, there was evidently a vital clue missing.

Dirac then recognized that Pauli's exclusion principle could help us to get over the difficulty. An electron could jump to a place on the lower curve only if this place was not already occupied by two electrons with opposite spins. What if all these states were normally occupied? This bold suggestion, of course, implies that empty space is full of such electrons in negative-energy states and, in fact, their density should even be infinitely large since these states extended over an infinite range of momenta. We might therefore expect an infinite density of negative electric charge, which would not make sense. But suppose that this negative charge density is in some way compensated by a positive charge density due to some other cause, so that there is still no electric field in empty space, in spite of all these hypothetical electrons, and that real charge is caused only by deviations from this situation. Then the presence of these electrons in negative-energy states would not easily make itself felt. We have seen that there existed in principle the possibility of an electron jumping from the upper to the lower curve, if there was a vacant state, and if the energy and momentum could be disposed of. It should equally be possible for one of these hypothetical electrons to jump from the lower to the upper curve, which is normally vacant, if it could receive the necessary energy and momentum.

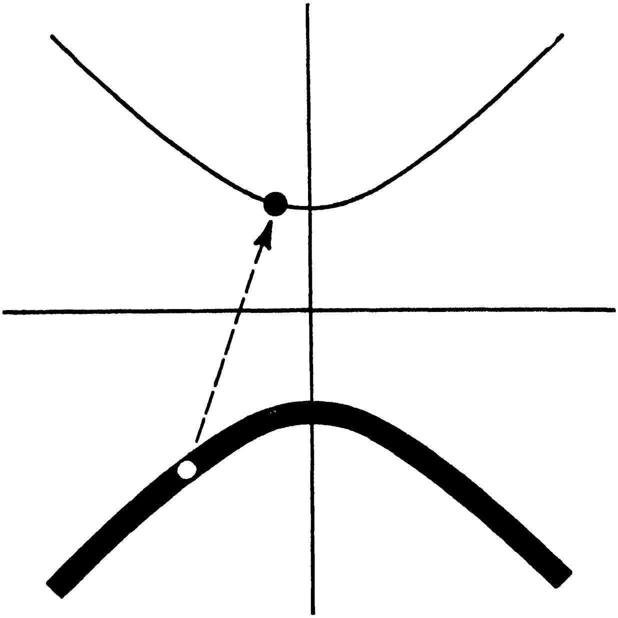


Fig. 61. Pair creation.

Just as the downward jump could proceed with the emission of two photons, an electron could be raised by the absorption of two photons. But in ordinary circumstances we never have sufficiently many light quanta of high energy present to have a reasonable chance of two of them being absorbed together. Another mechanism is for one light quantum to pass close by a heavy atom, with its strong electric field. In this case a single photon is sufficient to raise an electron. This has been indicated schematically in Figure 61. After the jump there is now a real electron on the upper curve and also a vacancy on the lower curve. What does this vacancy mean? Since the negative charge of all electrons on the lower curve was assumed to be compensated, the absence of one of them would show itself as the absence of negative charge and that is the same thing as positive charge. The absence of a particle with negative energy means that the energy has been increased and if the vacancy or "hole" is in a state with larger momentum so that the electron which originally was there has a large negative energy, then more energy had to be invested in creating this vacancy. The energy of a hole increases therefore as its momentum increases.

We see therefore that these holes behave in all respects like electrons with positive energy and with positive electric charge. If this picture is right, we must therefore expect that there should exist in nature, besides ordinary electrons, also positive electrons which, except for the sign of their charge, behave in every way like electrons.

We have seen that a photon passing by a heavy atom should be able to produce the transition sketched in Figure 61, i.e. to produce a pair of electrons, one positive, one negative. The least energy which the light quantum must have for this purpose is equal to the jump from the highest point on the lower curve to the lowest point on the upper curve and this is $2mc^2$ or about 1 MeV.

Conversely, if an electron and a positive electron meet, which in terms of our picture means really an electron on the upper curve and a hole on the lower, then the electron is able to fall into this hole and we have returned to the normal state of empty space. This can proceed either near an atom with the emission of one photon or, if the electron and hole meet in free space, with the emission of two photons. If the two electrons are moving very slowly the two photons have opposite and equal momentum, i.e. come out in in opposite directions, and each having the same energy, of about half a million electron volts.

These ideas were confirmed when the positive electron was discovered in the laboratory. If gamma rays of more than 1 MeV energy pass through a thin foil of, say, lead, in an evacuated tube, electrons are seen to emerge from the foil, some of which are deflected by a magnetic field in one direction, some in the opposite direction, so that it follows that some have a positive electric charge. If the positive ones are stopped in matter, gamma-rays are found to emerge from the place where they stopped and these gamma-rays are found to have energies of about ½ MeV each. Modern counters which respond to single photons can be used to prove that there are always two light quanta coming out in opposite directions at the same time.

Taken by themselves, these positive electrons or "positrons", as they are now called, behave just like negative electrons, but they do not live long in matter, because there are so many negative electrons present with which they can combine and be annihilated, their energy going into radiation. It is therefore not surprising that they are not normally found except where there are high-energy radiations capable of producing them.

In fact, if we imagined a state of affairs in which there were both positive and negative electrons present in large numbers but with a slight excess of the one kind, then whenever two of the opposite kind meet they would be annihilated and replaced by radiation until finally only electrons of one sign would be left, an extreme case of majority rule.

One interesting aspect of the creation and annihilation of electron pairs is that it confirms the idea of Einstein about the equivalence of mass and energy, in the sense that in this case the rest mass of the particles is also transformed into another form of energy.

If a slowly moving positron meets an electron, the two may, under the influence of their electric attraction, start revolving about each other before their annihilation takes place. During this time they then form an atom like that of hydrogen, except that the proton is replaced by a positron which has the same mass as the electron. Instead of the electron revolving around the heavier particle, the two will revolve about their centre of gravity. This combination which, of course, can exist for only a very short time, is called an atom of "positronium". In spite of its short life, skilful experiments succeeded not merely in proving its existence and measuring its lifetime but even in finding its energy levels in considerable detail.

In spite of all the successes of the "hole" theory of positrons, it might be objected that the picture of an infinite density of electrons in negative energy states, from which it starts, is not very satisfactory. The answer to this is that this picture is used only to help us look at the problem in the right way. When we describe in detail a situation involving electrons and positrons, we do not, in fact, have to introduce such electrons in negative states at all, but we specify only how many "holes" or positrons there are and how they move. The important thing is that we have laws to describe these particles, which do not predict any transition of an electron or positron into a state of the unreasonable behaviour which we discussed earlier on. Instead, the laws predict the possibility of pair creation and annihilation. Without these pair processes, it would not have been possible to remove the unreasonable kind of transitions, and the simple picture of Figure 61 is helpful in explaining how this comes about.

FORCES AT HIGH SPEED. MORE DIFFICULTIES

We see therefore how the fourfold wave of Dirac can account for the motion of free electrons or positrons and for the motion of either in a static field of force. But we meet new problems when we try to account for the interaction between electrons moving nearly as fast as light. It is then important to remember that the electromagnetic field which transmits the forces from one particle to another carries disturbances only with light velocity and, if the particles move quickly, one will experience not the force from the other where it is now, but where it was slightly previously, just as in watching a fast plane overhead we hear the sound come, not from where the plane is now but from a point further back, where it was when it emitted the sound that reaches us now.

This delay is known as the "retardation" of the force, and in order to allow for it in the quantum laws we must apply quantum theory also to the electromagnetic field, i.e. we must describe it in terms of photons.

The force of one charged particle on another then comes about in this way: one of the particles may emit one or more photons which may be absorbed by the other. Now to emit photons requires energy and the particles may not have much energy to spare, as for example, when they are nearly at rest. But here the remark applies which we made at the end of Chapter 8: quantum mechanics makes it possible to "borrow" energy for short times or over short distances. Thus if the two charges are close together then it is possible to borrow enough energy to produce photons of long or short wavelengths because, as the photon is going to be reabsorbed very near by, the loan can be repaid at once. On the other hand, if the particles are further apart, only smaller amounts of energy can be borrowed and only photons of longer wavelength will play a part. One speaks of "virtual" photons, to distinguish these particles existing "on credit", from real ones, which can get away.

In this way, the quantum theory of the electromagnetic field can be used to derive Coulomb's law between stationary charges, but the same reasoning can be applied to the more complicated case of charges in rapid motion. The laws about the "virtual" emission and absorption of photons which are required for this purpose are exactly the same as those needed in dealing with the real emission and absorption of light by atoms which we mentioned previously.

At this point a new difficulty appears. Even a single particle may emit and reabsorb photons. Since there is no question of a distance, the question of a minimum time for the loan does not arise, and therefore there is, in fact, no limit to the energy of the photons that a particle can emit and reabsorb. If all these possibilities are counted together the result would mean an infinite change in the energy of the particle, usually described as its "self-energy".

This trouble is not unexpected. Even from the point of view of Chapter 2, where we had seen that the energy density of the electric field is proportional to the square of its strength, the energy of the electric field near a point charge would be infinitely large. In fact, if a charge e is spread over a small sphere of radius a, one can work out that the total energy contained in the electric field is $\frac{3e^2}{5a}$. If we make the size of the sphere smaller, i.e. make a very small, the electric energy then becomes extremely large. It is therefore not surprising that the theory should predict an infinite energy in the field surrounding a particle whose charge is concentrated on a point.

This difficulty was known long before quantum mechanics; and the most natural way of avoiding it would be to consider the electron not as a mathematical point, but as having a finite size. Such ideas were discussed in particular by H. A. Lorentz, famous for his work which prepared the ground for the theory of relativity, as well as for the study of the behaviour of electrons.

However, this view of an extended electron is also difficult to maintain. This is because relativity does not admit the possibility of perfectly rigid bodies. The reason for this is simple to see. Suppose I strike a rigid body a sharp blow at one end. Then, since it cannot yield, its other end must start moving at once, and therefore the blow is transmitted from one end to the other instantaneously. But if this process is viewed by a different observer moving at high speed, then, on the laws of the Lorentz transformation of Chapter 6, the far end would start moving before the blow had occurred so that the effect would appear to precede the cause, which is unreasonable. In fact, it is essential to the ideas of relativity that it is not possible for any disturbance to be transmitted with a speed greater than that of light. On the other hand, if the near end of the body can move before the far end has started then the body must be at least slightly compressible. This is the case for all known solid bodies, e.g. for a bar of steel. If we hit one end of a steel bar with a hammer, the bar will yield slightly and it will take some little time before this compression travels to the other end of the bar. The speed with which this goes on is the velocity of sound in the steel, which is of the order of a few thousand metres per second. This is very fast for practical purposes, but still extremely small compared to the velocity of light.

This consideration explains why it is difficult to describe the behaviour of absolutely rigid bodies without violating the principle of relativity. All more recent attempts to regard the electron as rigid but of finite extent have failed, though, even to-day, few people would feel sure that some reasonable possibility might not have been overlooked.

An obvious alternative would be to assume the electron, in effect, compressible. Any body which is compressible is capable of internal vibration. In the case of a steel bar these are brought in evidence by hitting it with a hammer, when it will produce a ringing tone, which is always the result of vibration. Similarly, if an electron is hit, as in a collision with another particle, one should then expect that it can be set into vibration and therefore be capable of carrying energy besides the kinetic energy of its motion. This would show up in experiments through the existence of excited states of the electron, just as the possibility of internal motion in an atom is responsible for excited states of atoms. In the extensive study of collisions involving electrons at very high energies, no evidence of such excited states has ever been found. It might be that they exist at even higher energies than we have so far tried. But as long as we know nothing about the nature of such excited states their introduction would be too speculative.

In fact, since none of the experiments dealing with electrons so far has shown up any features indicating any kind of internal structure of an electron, it remains very doubtful whether the question of the internal structure of the electron has any sensible meaning, i.e. will ever be associated with any actual observation of its behaviour.

NEW METHODS. LAMB SHIFT

For a long time it seemed as if no further progress in the theory of the electron could be made while these questions remained unanswered, since any detailed description of the interaction of electrons through the electromagnetic field always led to infinite answers, connected with the fact that the energy content of the field surrounding a point charge is infinitely large. In the late forties there came, however, a new development which showed that, in spite of this difficulty, further information could be obtained. To understand this, consider for a moment again the case of a small but finite electron, in spite of the fact that, as I have pointed out, such a picture cannot be reconciled with the requirements of relativity. We should then expect that such an electron, if at rest, would be surrounded by its electromagnetic field and we would therefore have to add the field energy to the rest energy m_0c^2 of the electron, where m_0 is the rest mass of the electron. Because of the equivalence between mass and energy, increasing the energy is the same thing as increasing the mass, and therefore the electron and field together will have a mass $m_1 = m_0 + m'$, where m' is the field energy divided by c^2.

Now the important thing is that whenever we see an electron it is accompanied by its field and therefore the total mass m_1 must be equated with its actual observed mass. This real mass m_1 is measurable and is known, whereas the separate parts m_0 and m' are not known and may not even have a well-defined physical meaning. If we now reduce the radius of the electron, the field part m' increases, as we have seen. Since the total, m_1, is known, this means that m_0 must be assumed to be smaller, the smaller the electron radius. If we assume the radius very small, the electromagnetic mass m' may exceed the actual mass of a real electron and then m_0, the "mechanical" mass would have to be assumed negative. A particle of negative mass by itself is a very queer thing, as we saw before, but provided it is always accompanied by an electromagnetic field so that the overall mass is positive, no unreasonable consequences need arise.

We therefore try to describe the behaviour of the electron, not in terms of its mechanical mass m_0 and electromagnetic mass m', but in terms of the total mass m_1. It might then be hoped that the results will remain valid even if we go to the extreme and consider a point charge, although for this m' has to be infinite, and therefore m_0 infinitely negative.

Remember that the electromagnetic mass, m', is connected with the energy of the electromagnetic field around the electron, which, as we have seen, is described in quantum mechanics by the possibility of the electron "borrowing" energy to emit light quanta for short times and reabsorbing them. Now consider such an electron moving not freely, but in the neighbourhood of a centre of force, for example, in a hydrogen atom. Then, in the time between the emission and reabsorption of a photon, the electron may have changed its velocity under the influence of the applied force. The process of emission and reabsorption of photons, and therefore the electromagnetic mass m', is therefore changed by the presence of the forces. But this change will concern mainly photons of long wavelength because short-wave photons require a high energy and, by the uncertainty relation, can therefore exist only for a very short time; in such a short time the external force cannot have accelerated the electron much. While the total electromagnetic mass m' is infinitely large for a point electron because of the effect of very short-wave photons, the amount by which it is changed in the neighbourhood of a centre of force is consequently finite even for a point electron.

From such considerations follows a small correction to the energy levels of the hydrogen atom. This differs from one level to another, because it depends on the probability of the electron coming close to the proton, which is different for the different states.

It is interesting to note that this whole development did not start as an abstract speculation but was stimulated by the discovery of unexplained anomalies in the hydrogen levels. These had been measured by means of a new experimental technique developed by Lamb and his collaborators at Columbia University, New York.

To appreciate the delicacy of the problem from the experimental point of view, we must refer to the energy levels of the hydrogen atom as, for example, in Figure 49. In that context, we discussed the solutions of the non-relativistic Schrödinger equation, in which the energy just depended on the number, n, which was 1 for the ground state and 2 for the first excited state. For $n = 2$, there are, as we had seen, two different kinds of motion of the electron, one with no angular momentum and the other with an angular momentum of one unit. In hydrogen, these should have exactly the same energy, which lies about 10 electron volts higher than that of the ground state. This treatment does not take account of the spin of the electron. Actually, because of the spin, the energy of the state with one unit of angular momentum is slightly different according to whether the spinning motion of the electron is in the same sense as its revolution in the orbit, or in the opposite sense. The difference, the so-called fine structure of the hydrogen levels, can be calculated from the equation of Dirac since, as we had seen earlier in this chapter, this gives a natural treatment of the spin. One finds a distance of about 1/200,000 of an electron volt. In the past, this difference had been detected through the presence of two closely adjacent lines in the hydrogen spectrum. It is clear that very careful spectroscopic work is necessary to show up, and still more to measure, the small separation between these lines.

The new method consisted in observing directly the transition between two such closely adjacent levels. Since the energy difference is then so small, the frequency of the light quanta absorbed in the transition is correspondingly small. We are here, in fact, not dealing with light in the ordinary sense but with short radio waves of a wavelength of the order of 20 cm. Equipment for producing such "microwaves" had been developed intensely during the war, because of its connection with radar work. It would take us too far to review the experiment in detail. It was possible to induce, by means of such radio waves, the transition between the two hydrogen levels. The frequency for which this takes place then gives directly the energy of the photons causing the transition and therefore the energy difference. This did not agree with what one had expected, since the two states concerned should have had exactly the same energy, and physicists began to wonder about the cause of this discrepancy.

This contradiction led to a renewed study of the way in which the electric field of an electron reacts back on its motion, and it was largely as a result of this stimulus that the new methods, which I have described, were developed. By these new methods it was then possible to calculate exactly what the energy difference should be and the result agrees with the most accurate experiments to a few parts in 10,000, a remarkable tribute to both experiment and calculation.

A similar subtle effect was discovered soon afterwards: the magnetic field of a spinning electron was found to differ by about three parts in a thousand from what the simple Dirac wave equation required, and this also was explained by Schwinger in terms of the virtual production and absorption of photons.

The very exact confirmation of the theory in these and other tests suggest that our present approach to the problem of an electron with its own field is on the right lines. In spite of this it appears probable that at least the form in which these laws have so far been expressed is not yet final because, as we have seen, they would seem to predict infinite results for the self-energy of an electron, to be cancelled only by assuming an infinitely negative mechanical mass. It may be that any such division into mechanical and electromagnetic mass has no meaning and that the same laws can be expressed in a form in which the two parts of the mass do not appear. One might imagine that such a reformulation could still lead to exactly the same answers as the present procedure. But this remains to be seen.

It is important to stress that the complications arising from the interaction of an electron with its own field concern only very small corrections which need experiments of high accuracy to show up at all. The reason for this is that the importance of this self-field is governed by the number $\frac{2\pi e^2}{hc}$ (where e is the electron charge, h Planck's constant, and c the light velocity), which we had already seen to be a very small number, $\frac{1}{137}$. For example, this number comes in when we estimate the probability of light emission in an excited atom.

It turns out that the probability of the emission of a photon during a time as long as a period of revolution of the electron in the excited state, is about $\frac{1}{137}\left(\frac{v}{c}\right)^2$, where v is the velocity of the electron. We saw that, for hydrogen, and for electrons of similar energy in other atoms, $\frac{v}{c}$ is again about $\frac{1}{137}$. Hence, the electron has in each revolution only one chance in 137^3 of emitting a photon; it will take, on the average, about 137^3 revolutions before it drops by light emission to a lower state. This means that we may treat the motion of the electron to a very good approximation as if the possibility of emitting light did not exist. The same number occurs again in the question of virtual emission and absorption of photons so that, in the first place, the corrections due to such emission and absorption are altogether small. Furthermore, in discussing these corrections, we may disregard the chance of an electron emitting a second photon before the first one has been reabsorbed as even smaller and thus consider only one photon at a time. Corrections due to the simultaneous presence of two or more photons are at most 1 per cent of those due to one photon.

The fact that the "fine structure constant" $\frac{2\pi e^2}{hc}$ is so small therefore helps us very greatly in the study of electron problems. In Chapter 11 we shall meet a similar situation in which the corresponding number, however, is not small, and this gives rise to difficulties which, at the present, are seriously holding up the development of the subject.

10

The Atomic Nucleus

RADIOACTIVITY. ALPHA-DECAY

In all the preceding chapters our aim has been to find the laws of nature which govern the behaviour of objects and of substances which we meet in practical life. The search for these laws has shown us the existence of atoms and has led to the study of their internal structure and their behaviour. The laws of quantum mechanics, together with the older laws of mechanics and of electricity and magnetism, made it possible to deduce all the characteristics of atoms from the knowledge that each atom contains a nucleus with a certain positive electric charge and that this would normally be compensated by the negative charge of electrons surrounding it. In this way it became possible to describe atoms, and therefore also matter made of atoms, in almost every detail. This allows us, not merely to understand what happens in ordinary matter, but also to use this knowledge to utilize matter for practical ends.

We can now understand why some substances are heavy and others are light, why some substances corrode easily under the action of chemicals and some are stable, some are hard and others are soft. We know why and how metals conduct electricity and why some substances are good conductors of heat, why some salts are coloured and others glow in various colours when exposed to the invisible ultra-violet light.

However, some questions remain unsolved. The most obvious of these concerns the nucleus of the atom. What are the differences between the nuclei of the 92 or so different kinds of elements known in nature? How can we account for their differences in weight, i.e. why is the nucleus of oxygen, and hence also the atom of oxygen, about 16 times heavier than that of hydrogen, or why does a lead atom weigh over 200 times as much as one of hydrogen?

These questions lead us naturally to a study of what happens inside the atomic nucleus. They form part of our search for the basic laws of nature, and they have to be answered before we can claim that our description of the behaviour of matter is complete. However, in general, they will not help to tell us how, in ordinary circumstances, matter behaves. It might be thought that they will not aid us in using the laws of nature for practical purposes since, in all the practical designs of the physicist, the chemist or the engineer, the atomic nucleus plays merely the part of a fixed and unchangeable piece which after any chemical process or change of temperature, after any mechanical treatment or any electric or magnetic change, comes out to be just the same as it was before.

The study of the atomic nucleus was therefore undertaken purely to satisfy intellectual curiosity but, as every reader will know, it turned out that there were special circumstances in which it was possible to release large amounts of energy from atomic nuclei. Through the discovery of what is now called "atomic energy" the knowledge of nuclear structure turned out, after all, to have far-reaching practical consequences.

In spite of this exception it is true that, in the study of the nucleus, we are talking about facts which do not normally manifest themselves through the effect they have on our everyday experience. In talking about atoms we could illustrate our conclusions in terms of the melting of ice or the boiling of water, the action of a magnet, or the colour of a neon tube. The problems of the nucleus arise only in the experience of the physicist with his specialized devices for probing the nucleus. The discovery of atomic energy does not really change this position since the working of an atomic power station will not in the immediate future and the detonation of an atomic bomb will never, we hope, be part of our common experience.

It will therefore be more necessary than in previous chapters to give some indication of the actual experiments from which our ideas of the nucleus are derived, and to refer to the devices for producing and detecting very fast particles, whose rapid development in the last twenty-five years had made the study of the nucleus possible.

The beginnings of the physics of the nucleus are linked with radioactivity. We have already referred, in Chapter 4, to radium, uranium and other elements extracted from minerals, which send out radiations of three kinds, in particular the "alpha-rays" which Rutherford used in his work on the structure of the atom .We saw that the alpha-rays were beams of positively charged particles whose mass and charge were the same as that of the helium nucleus. Where do these come from? The chemical purification of the radium which is used makes it sure that there was no helium mixed with it; the only place where such a heavy particle of positive charge could exist without altering the chemical properties of the radium atom is inside the nucleus itself. This was confirmed by experiments which showed that chemical changes, which affect the motion of the outer atomic electrons, could not speed up, retard, or otherwise alter the alpha-activity.

As soon as our present ideas about the structure of the atom emerged, it became clear therefore that what we were witnessing here was a break-up of the nucleus of the atom, in which one part was thrown out with very high speed. Careful observation showed, in fact, that the activity, i.e. the strength of the radiation of a given sample of radium or any other radioactive element, diminishes with time. In the case of radium, this is rather slow and the amount of alpha-radiation diminishes to one half only after about 1,600 years, but, for example, for the so-called radium emanation, a gas which is also found together with radium, for a reason that we shall see presently, the activity goes down to one half after about four days. Suppose then that we pump off a certain quantity of emanation from some radium, seal it in a thin-walled tube which allows the alpha-particles to pass and watch the rays. After about four days, there will only be half as many rays coming out as at the beginning but, otherwise, they are in no way distinguishable from those we get in the first hour. The only way to explain this is to say that all the emanation nuclei are the same and indistinguishable, but each of them is liable to break up at random and, on the average, this happens to only about two in every million such nuclei in every second.

We might picture this in the following way. Suppose we made up a large number of little boxes, each of them containing a large number of compartments linked by complicated passages and only one exit, so that each formed a maze. In each of them we place an insect which is supposed to crawl around without knowing the way, and without being able to benefit from past experience. Occasionally, by chance, one of them would find the way out and escape, but those which had not yet found the exit would still be wandering around, and their chance of escape would be no greater and no less than it was at the start. If, after a few days, one half of them had got out, the remaining half would be in the same condition as at the start, and after the lapse of the same period one half of them will have escaped, thus leaving only one quarter of the boxes still occupied, and so on. Similarly also our quantity of radium emanation will diminish to one half after the first interval of about four days; waiting again for the same length of time, it will be down to one quarter, and then to one eighth and then to one sixteenth of the amount we started with, and so on. What has become of the nuclei from which an alpha-particle has been thrown out? We know that electric charge and mass are conserved. Since the alpha-particle carries a mass of four times the mass of a proton and twice its charge, what remains must be a nucleus which has a charge less by two units than that of emanation and whose mass is less by four units. Such a nucleus will form an atom which, instead of the 86 electrons of radium emanation, contains only 84, and this gives it different chemical properties. We have therefore witnessed the transformation of one chemical element into another. This is confirmed by chemical tests.

This means we must abandon completely the idea of atoms as indivisible and immutable objects. We have already seen in previous chapters that this idea did not apply to the atom as a whole. We have not previously met changes in the nucleus itself; we could alter the number of electrons around a nucleus, but then the remaining electric charge would ultimately lead to a restoration of the correct number and the atom would return to its natural form. Here, however, we are concerned with a permanent change in the nature of the nucleus, and, consequently, in the kind of atom that it will form.

In many cases, the nucleus remaining after the radioactive process is itself radioactive, i.e. it will in due course emit another alpha-particle or some other radiation of a type we shall meet later. This is noticed by the fact that if, for example, we start with very pure radium and pump off all the emanation that might accompany it, after a time, some more emanation is formed which shows up by the alpha-particles it emits and which can again be pumped off as a gas. We conclude therefore that radium emanation is the "decay product" of radium, i.e. its nucleus is what remains after a radium nucleus has lost its alpha-particle. The decay product of the emanation is again radioactive and in this manner three "families" of radioactive elements have been identified; in each of these one member is the decay product of the preceding one.

The speeds with which the alpha-particles come out in these reactions are very large. It is customary to measure their kinetic energies in million electron volts (MeV). One MeV is the energy which an electron or a particle with the electric charge of an electron, acquires in passing through a potential difference of one million volts. By comparison, the energies of the outer electrons in atoms are, as we saw, a few electron volts. The energies of alpha-particles from radio-active substances are a few MeV; for example, in the case of radium just under 5 MeV. Remembering that the kinetic energy is $\frac{1}{2}mv^2$, where m is the mass and v the velocity, we see that even though the mass of an alpha-particle is over 7,000 times that of an electron, its speed is over ten times that of the electron in a hydrogen atom. We are therefore dealing with very large energies on the atomic scale. We should not, of course, forget that for practical purposes these are still minute amounts of energy, since we are not accustomed to handling single atoms. If we could influence the decay of radium in such a manner as to make all the radium atoms in, say, 20 gm. of radium decay at the same time, this would release a million times the energy of an ordinary chemical process involving the same number of atoms, such as the combustion of 1 gm. of coal. But this is an idle speculation, since nothing we can do will accelerate the decay of radium, and therefore this energy will not come out at once but only over a period of many thousands of years. With a short-lived substance like emanation, the energy is produced in a more reasonable time, but just for that reason it is not practicable to accumulate anything like grammes of emanation because it is formed only extremely slowly from radium and decays almost at once after being formed.

WAVE MECHANICS SOLVES ANOTHER PARADOX

In a sense, the energies of the alpha particles are not as high as one might have expected, and in the early work of Rutherford this appeared as a difficulty. If the alpha-particle comes from the nucleus, then from the start it is subject to the electric repulsion of like charges, which will accelerate it. If we know the distance from which it has started, the conservation of energy can tell us how much kinetic energy it must have built up. Now according to Coulomb's law, the potential at a distance, r, from a charge of Z electron charges is $\frac{Ze^2}{r}$ and the potential energy of an alpha-particle, which itself carries a positive charge of $2e$, should therefore be $\frac{2Ze^2}{r}$. And this must equal the kinetic energy which the alpha-particle has acquired on the way. Since we know the energy with which it ultimately comes out, it follows that it could not have started its journey from a distance, r, closer than about 6×10^{-12} cm., provided that even at such close approach, the force between the nucleus and the particle still obeys Coulomb's law. Now the original experiments of Rutherford on the deflection of alpha-particles in passing through matter had shown that Coulomb's law is, in fact, valid to much closer distances than that, and from modern improved versions of these experiments, we know to-day that the radius of even the largest nucleus is only about 1×10^{-12} cm., in other words, about 6 times smaller than the distance from which the alpha-particle ought to have started at rest to come out with the velocity with which it does come out.

This is about as surprising as if we found a golf ball dropping at our feet with a gentle thud and looking up for its origin saw nothing above us other than a bridge 100 ft. high. There would seem to be no reasonable place less than 100 ft. high from which it could have started, and if it had been dropped from the bridge it should have landed with a much more violent impact.

This difficulty remained unresolved until the laws of wave mechanics were discovered, and it was then shown, by Gamow, and by Condon and Gurney, to be just an example of the passage through a potential barrier which was discussed at the end of Chapter 8. The position is illustrated in Figure 62, which shows the potential energy of an alpha-particle against its distance from the nucleus. At distances not too close to the nucleus, the only force is the electric repulsion, and therefore the main part of the curve represents the effect of the Coulomb law. The potential energy rises towards the nucleus because in moving a charge close to the nucleus we must overcome the electric repulsion. At very close distances, other forces will come into play, which we shall discuss later; these are attractive. For this reason, the curve has been shown to fall again steeply. Its details inside the nucleus are not important for the moment. Now if the actual energy of the alpha-particle corresponds to the broken line in the figure, *A* again denotes the point from which the particle will have to start at rest to be thrown out by

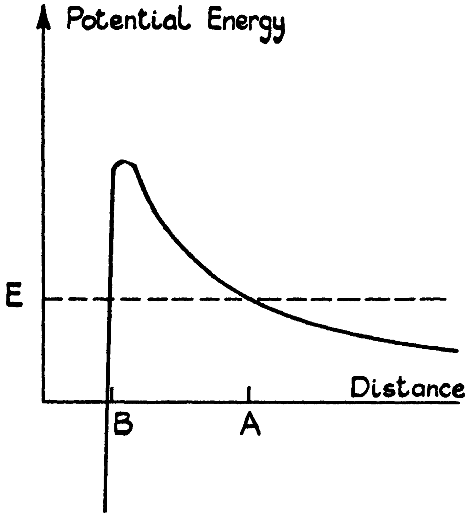


Fig. 62. Alpha-decay.

the repulsive force with just the right kinetic energy. It could not have come from anywhere between *A* and *B* but left of *B* there is also a region where it could move with the given energy. It will be seen that this situation is, in principle, the same as that in Figure 55 and that, therefore, we have a situation in which quantum mechanics predicts a small chance of escape through the potential barrier of a particle which starts inside, i.e. to the left of *B*, with the energy shown.

This argument therefore explains, not only that particles with such a relatively low energy can get away from a small nucleus, but also that the probability of their doing so may be very small if the potential barrier is high and wide enough. This therefore explains why, in a short time, only an extremely small fraction of the nuclei will undergo the decay. In our comparison with the insect in a maze, the potential barrier now takes the place of the maze. It is evident that the chance of escape is smaller for lower energy of the alpha-particle, because if we lower the broken lines in Figure 62, the height of the barrier, i.e. the height of the full line above the broken one, increases and also with it the width of the barrier, i.e. the distance from *B* to *A*. It is easy to calculate the half-life, i.e. the time we have to wait before half the atoms have decayed, and how it varies with the available energy. Although the half-lives range from millions of years to small fractions of a second, the calculations agree well with observation, thus again confirming one of the important features of quantum mechanics and, at the same time, showing that the interpretation of alpha-decay is along the right lines.

SPLITTING THE NUCLEUS

Up to this stage, the part played by the nuclear physicist is somewhat like that of an astronomer who can observe and record the behaviour of stars and planets, but who has to take them as they are, without being able to influence their fate. Rutherford set himself the aim of breaking up a nucleus artificially.

He succeeded in this by watching carefully the passage of alpha-rays through matter, for example, through air. We have already seen that, most of the time, these pass merely through the outer parts of the atoms in which they are not appreciably deflected from their straight path. Occasionally one may come close enough to a nucleus to be strongly deflected by the electric force. On very rare occasions one will make a direct hit on a nucleus. This is rare because we are firing blind at a target with an extremely small bull's eye. We remarked previously that the diameter of even the largest nucleus is about 10,000 times smaller than that of an atom, and the area which the nucleus exposes to the projectile is therefore the square of 10,000 or a hundred million times smaller than the atom. Hence, on passing an atom, a particle has only one chance in a hundred million of scoring a direct hit.

Now it is true that, in the course of its passage, the alpha particle will pass through very many atoms and therefore it has effectively, not one try, but many. But this does not go on indefinitely because, in passing through the outer parts of atoms, the alpha-particles pass close to electrons and disturb them. This results in a kind of friction and therefore the particle is being slowed down all the time. After a certain number of such passages, it comes to rest, and even before it stops its speed is no longer sufficient to overcome the repulsive force which acts against it approaching the nucleus. The particle passes roughly through 10,000 atoms before it comes to rest, and this improves its chance of scoring a bull's eye. Its chance of making a direct hit at least once is of the order of one in a hundred thousand to one in a million, according to the speed with which it starts and the kind of matter through which it passes. When it does make a hit, for example, in a nitrogen atom in the air, the result is that the alpha-particle gets stuck in the nucleus and a proton (hydrogen nucleus) is thrown out. This proton has a much higher speed than the original alpha-particle and therefore will travel further before stopping. This is the reason why it was possible to detect these few protons in the presence of so many alpha-particles, because they penetrate to distances beyond the end of the range of all the alpha-particle tracks.

Rutherford's work was, at that time, done with the aid of a scintillation screen. When alpha-particles from a radioactive substance fall on a screen covered with a suitable salt, they cause it to glow, a phenomenon which is utilized in the luminous figures and hands of a watch. If the alpha-rays are numerous, this is just a continuous glow, but if one uses a very weak source, the light becomes irregular and one can, in fact, detect the weak glow or "scintillation" caused by a single alpha-particle.

If a thin metal foil is placed between the source and the screen, one can make the scintillations stop completely because all the alpha-particles are stopped in the foil. In a heavy metal, none of the alpha-particles can score a direct hit because of the strong repulsive forces from the nuclei. Now if the space between the source and the foil contains air or pure nitrogen a few scintillations will again be seen. These are due to protons which have a longer range than the original alpha-particles and therefore are not completely stopped by the foil.

The conclusion to which Rutherford came was that each proton was a fragment of the nucleus which had been hit. On capturing an alpha-particle, with its two units of positive charge, the charge on the nucleus had been increased by two, but the ejection of a proton took away one charge so that in the result the charge on the nucleus had increased by one, thus turning the nitrogen nucleus into one of oxygen.

At that time work in this field was slow and laborious because of two limitations. The sources of radiation available for bombardment were limited, since one had to rely on the radio-active radiations provided by nature. The detection of the results was difficult because the use of a scintillation screen required the observer to spend long hours in complete darkness watching faint flickers of light which were only just detectable by a patient observer. Nuclear physics was revolutionized, therefore, by developments which have given us very much more powerful beams of particles and very much more convenient and sensitive methods of detection.

NEW EQUIPMENT

One of the most spectacular methods of detection consists in the cloud chamber invented by C. T. R. Wilson. In this, water vapour is made to condense into little drops on the ions which a fast charged particle forms by knocking electrons out of atoms. This results in a trail very similar to the familiar vapour trails produced in suitable weather conditions by high-flying aircraft. Figure 63 is a diagram of the tracks in such a cloud chamber photograph. A source of alpha-particles is near the bottom of the picture and one sees the straight track of many alpha-particles. One of these in particular has made a collision with a nitrogen nucleus as described above. One sees the fainter trail of the resulting proton (fainter because it moves at a higher speed), and a heavy track shows the nucleus recoiling after the collision. Pictures of this type, in which it is possible to see a nuclear collision, were first obtained by Blackett. A similar method using photographic plates will be discussed later.

Another important method uses a gap between the plates of an electric condenser, with a suitable potential difference, so that the electrons, which are knocked out of atoms by the passage of a charge particle, are attracted by the positively charged plate. In this way a small pulse of current passes through the condenser, and this is then amplified by electronic devices and can then be recorded by purely mechanical means or in other ways. Such a device is known as a counter. The best-known forms of this are the Geiger

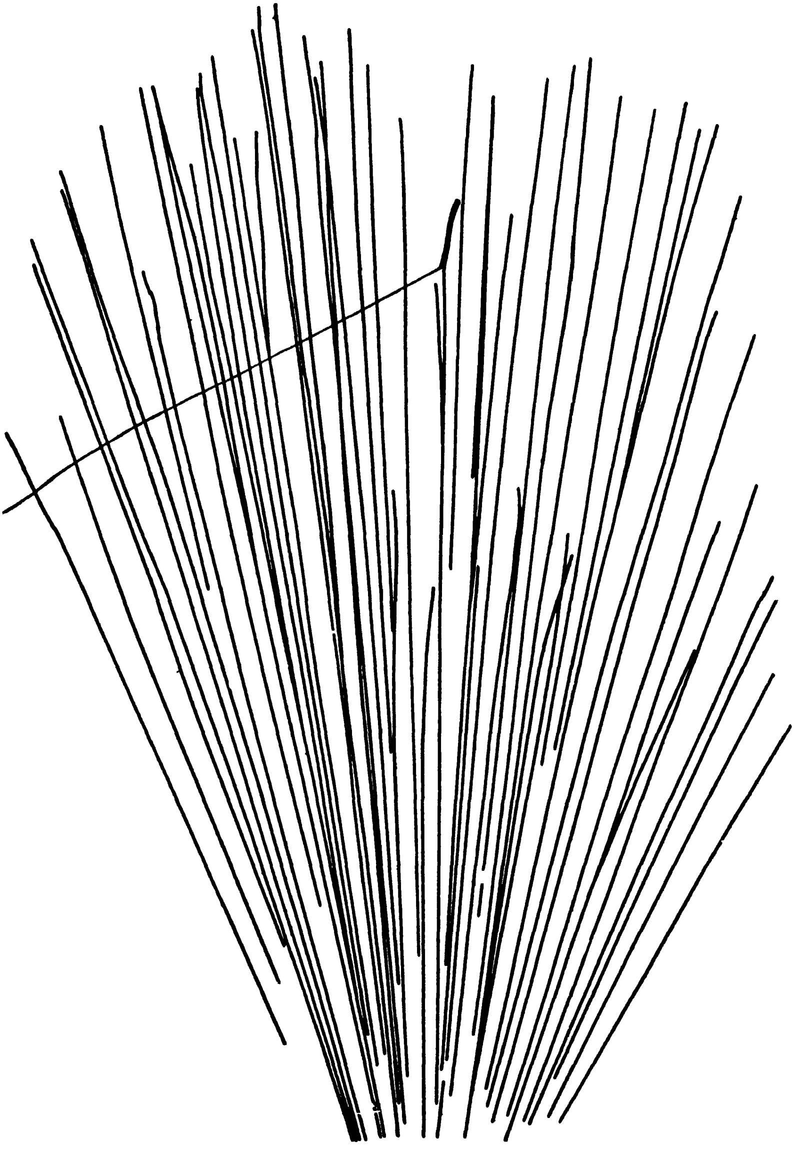


Fig. 63. Cloud chamber tracks of alpha particles, showing the ejection of a proton from a nitrogen nucleus.

counter and the Geiger-Müller tube in which the applied potential between a wire and a tube surrounding it is just not large enough to start an electric discharge. If a passing particle sets free a few electrons this may be sufficient to start a discharge, which is arranged to last only for a short time, so that the counter is ready to record the next particle. One of the latest devices to be developed is based on the old idea of the scintillation counter, but by combining the old scintillation screen, or alternatively a large transparent crystal, with a modern version of the photo-electric cell one can make the flash of light be detected electrically so that one need no longer rely on the eye of the trained observer; this makes the counting easier, faster and more reliable.

Together with these developments, great progress was made in artificially producing beams of fast particles. This was first achieved by Cockcroft and Walton, who designed an electric circuit in which, by means of transformers and rectifiers, a potential difference of a few hundred kilovolts (i.e. a few hundred thousand volts) was applied between the ends of an evacuated tube. At the positive end of this tube protons, or other nuclei, are stripped of their electrons by a small electric discharge and they are then accelerated by the electric field of the tube and leave it at the negative end with an energy of a few hundred MeV. With this apparatus, Cockcroft and Walton could disintegrate atomic nuclei without the aid of radioactive radiations.

An important variant of this is the Van de Graaff generator, which has a similar accelerating tube, but uses a different method for generating a constant voltage. One now finds in many nuclear physics laboratories high-tension sets either of the Cockcroft-Walton or Van de Graaff type, capable of giving particles of several MeV.

It would be difficult to attain higher energies by such a direct method, since a potential difference of a million volts is capable of producing a spark in air across a distance of several feet and therefore the length of a tube carrying such a potential difference must be large to avoid spark discharges which would make the attainment of large voltages impossible. One can get a little further by enclosing the whole apparatus in a tank filled with air or some other gas at high pressure, which reduces the spark formation, but it does not seem practical by this method to get beyond a few MeV.

This limitation is overcome by a number of machines which accelerate particles without using very high potential difference. Of these, the linear accelerator is easiest to describe. This is, in principle, a series of short metal tubes following each other as in Figure 64. An alternating voltage is applied to these tubes in such a manner that when the first is at a positive potential, the second is negative, the third again positive and so on. Consider then a positively charged particle (e.g. a proton) which passes from the first to the second tube when the first is positive. As it passes through the gap between them it finds itself in an accelerating field and its velocity will increase. The length of the tube is so arranged that the voltage is reversed by the time the particle has reached the end of the second tube, so that the second tube is now positive and the third negative. Thus the particle is again accelerated. If a particle passes in this way through ten gaps, the energy it has acquired is ten times that which it receives in a single gap, and therefore the particle acquires the same energy as it would get from passing through a potential difference ten times as large as that between two tubes. In principle, this device can be extended as far as one wishes, but the provision of electrical power for a large number of accelerating gaps is a complication.

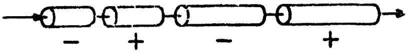


Fig. 64. Linear Accelerator.

This difficulty is overcome in the "cyclotron" invented by E. O. Lawrence. In this, an alternating electric voltage is applied to two open metal boxes placed facing each other with their open sides. The path of the charged particles is bent into a circle by applying a magnetic field. We have seen previously that in a magnetic field a charged particle is deflected from its straight path and describes a circle. The radius of the circle increases as the speed of the particle increases but, as long as relativity is unimportant, the time of revolution is the same for a slow particle which revolves in a small circle circle and a fast particle in a larger circle. The frequency of the alternating current is adjusted so that it goes through a complete cycle as the particles complete a revolution. In this way, one can arrange that the particles cross from the one box to the other always when the first is at a positive and the second at a negative potential and that they cross back when the potential difference is reversed. Thus they are accelerated twice in every revolution. By this device it is possible to bring particles to energies corresponding to many millions of volts while the alternating potential difference between the boxes is only a fraction of a million volts. For convenient work in nuclear physics, the cyclotron has energies ranging from a few to 20 or 30 MeV.

In recent years interest has shifted to still higher energies. Cyclotrons are now in existence which can accelerate protons to energies of 400 to 500 MeV. This is about half the rest energy of the proton and therefore relativity is no longer negligible. It is then no longer true that the period of revolution of such a particle is the same whatever its energy, but the period begins to increase appreciably, and in such high-energy machines a somewhat more complicated scheme has to be adopted. Instead of keeping the frequency of the alternating current constant one arranges for the frequency to decrease as a group of particles is being speeded up, so as to remain in step with their motion. Such a machine will accelerate only particles which start off at about the right time, so that it produces, not a continuous beam, but a series of "bursts" of particles. When one group of particles has been accelerated to the maximum energy, the frequency goes up again and the machine is ready for the next batch of particles.

With this principle, there is no limitation to the energy which may be reached, except that a particle of very high energy moves in a circle of very large radius, requiring therefore an extended magnetic field, i.e. a large magnet. These machines therefore become expensive pieces of heavy engineering. To keep down the size of the magnet, one may arrange to keep the particles moving on a fixed circle rather than starting them off on a very small circle which gradually grows. This is done by making the magnetic field also vary during the time that each batch of particles is being accelerated, starting off with a very weak field and making it grow as the particles get faster. In this case, one has to provide the magnetic field over only a ring-shaped region, in which the particle stays from beginning to end. Such a machine, in which both the magnetic field and the frequency are changing to keep in step with the acceleration of each batch of particles, is called a "synchrotron", and at the time of writing, particle energies of several thousand MeV have been reached by this method, while plans for new machines going up to 20 to 30,000 MeV are under discussion. However, such machines do not properly belong to the present chapter since their main use is for problems which will be studied in the next chapter.

NEUTRONS, ISOTOPES, MASS DEFECT

After this brief review of the equipment and methods of the nuclear physicist, let us return to the results of his work.

One very important step in the understanding of the structure of the nucleus was the discovery, in 1932, of the neutron by Chadwick. This is a particle of approximately the same mass as the proton (it is actually slightly heavier) but distinguished from the proton by carrying no electric charge. As its name indicates, it is electrically neutral. Neutrons are in general produced in the same way as the protons in Rutherford's first experiment, namely by the collision of some charged particle with the nucleus.

When such neutrons pass through matter they exert no force on the electrons and can therefore pass right through atoms without disturbing them at all and without losing speed on the way. They therefore go on until they make a direct hit on some atomic nucleus, and they can be detected only through the results of such collisions. This makes neutrons very convenient projectiles to use in bombarding nuclei, if one wants to study their behaviour. Whereas a charged particle, such as a proton or an alpha-particle, is normally stopped by friction before it has had a chance to make a direct collision with a nucleus, neutrons will always go on until such a collision takes place. However, since there exists no means of accelerating neutrons to high energies, the chief neutron sources available to the physicist (other than reactors to which we shall return later) rely on the bombardment of nuclei by fast charged particles. While we obtain an interesting collision for almost every neutron, a large number of charged particles must have been accelerated and wasted to produce this neutron in the first place.

The fact that neutrons and protons can be knocked out of nuclei as the result of collisions suggests that they are contained in every nucleus, and in fact we now know that nuclei are built entirely of neutrons and protons.

This explains at once a fact which has been noticed for a very long time, namely that the atomic weight of many elements were almost exactly multiples of the atomic weight of hydrogen. For example, the mass of an alpha-particle or helium nucleus is very nearly four times the proton mass. Since its electric charge is twice the proton charge, this suggests that it consists of two protons and two neutrons. Similarly, an oxygen nucleus, with 8 units of charge and 16 units of mass, consists of 8 protons and 8 neutrons.

Before this explanation can be accepted, one must understand why the atomic weight of some elements do not obey this rule. A typical exception is the case of chlorine with an atomic weight of about 35.5. The explanation for this had been given a long time ago by Soddy, who suggested that such elements were in fact mixtures of different kinds of atoms of different weights which were chemically indistinguishable. In other words, chlorine atoms are not all identical. Some have nuclei whose mass is 35 units and some have nuclei with 37 units of mass. Both have the same charge, viz. 17 times the proton charge, and therefore both kinds of atom contain 17 electrons, and this determines their chemical behaviour. Such atoms of the same chemical nature but different mass are known as "isotopes" from Greek words which indicate that they occupy the same place in the series of elements. This hypothesis was confirmed by means of the mass spectrograph invented by Aston. This is an instrument in which beams of ions are deflected by electric and magnetic fields. From the extent of their deflection it is then possible to determine their charge in relation to their mass, as we saw in Chapter 4. If an element such as chlorine is analysed in the mass spectrograph, two separate lines are seen on the receiving plate. One represents the mark due to atoms of mass 35, and the other is made by atoms of weight 37, which, being slightly heavier, are somewhat less strongly deflected from their straight path. Such work has shown that most elements are really mixtures of isotopes, but in many cases one of the isotopes is much more abundant than the others, so that the average atomic weight as measured by the chemist is very close to that of one particular isotope.

Since we know that atomic nuclei consist of neutrons and protons, we conclude that different isotopes of the same element must contain the same number of protons (in order to have the same electric charge) but different numbers of neutrons. For example, the two isotopes of chlorine which have been mentioned, contain 17 protons each and, respectively, 18 and 20 neutrons. Even the simplest element, hydrogen, was found to possess one isotope which occurs in nature. This is the so-called heavy hydrogen of atomic weight 2, and its nucleus evidently consists of one proton and one neutron. It is rare, making up only one part in five thousand of natural hydrogen. There is even a slight difference in chemical behaviour between ordinary and "heavy" hydrogen, and for this reason heavy hydrogen has been given a separate chemical name "deuterium", and its nucleus is called a deuteron.

We shall not go into the reason for this slight chemical difference, which has to do with quantum effects in the vibration of the hydrogen or deuterium atom inside a molecule.

We might now expect that the mass of a particular nucleus or the atomic weight of a pure isotope should equal exactly the combined mass of all the neutrons and protons which are contained in it. But this also is not exactly true. The atomic weight of each nucleus is found to be somewhat less than that of its constituent parts. This was verified both by accurate observations by the mass spectrograph, and also from comparison of the atomic weights as measured by the chemist, in those cases in which only one isotope of each element mattered. For example, whereas the atomic weight of oxygen ought to be more than 16 times that of hydrogen (since a neutron is slightly heavier than a proton), it is actually about 1 per cent less than this.

This "mass defect" was explained by the equivalence of mass and energy, which forms part of the laws of relativity (see Chapter 6). In forming a nucleus of oxygen by combining 8 protons and 8 neutrons, energy must be set free and, conversely, energy must be put in before we can split an oxygen nucleus into its parts. In other words, the nucleus contains less energy than 8 protons and 8 neutrons separated from each other, and since mass and energy are the same thing, the reduced energy must mean a reduced mass. We now know from nuclear collisions how much energy is required to knock a particle out of a nucleus, and how much energy is set free by the capture, for example, of a deuteron in a nitrogen nucleus to form a nucleus of oxygen.

Similarly, we can find out how much energy is set free by forming nitrogen from still lighter nuclei, and in this manner we can find the total "binding energy" of an oxygen nucleus to be about 120 MeV. If we divide this by the square of the light velocity we should find, according to Einstein's relation, the reduction in mass. Since 1 MeV is the energy equivalent to the mass of two electrons we see therefore that the "mass defect" should be about 250 electron masses or a little over 1 per cent of the mass of the oxygen atom. The comparison between mass defects and binding energies give excellent agreement in all cases, thus confirming again both our picture of nuclear composition and the principles of relativity.

NUCLEAR FORCES

The next fundamental problem that arises is that of the nature of the forces which hold the neutrons and protons in a nucleus together. Evidently, these are not electric forces because all the protons have like charges and should therefore repel rather than attract each other, and the neutrons carry no electric charge at all.

It is also clear that the forces must be very much stronger than electric forces. This can be seen from the fact that a nucleus is roughly 100,000 times smaller in diameter than an atom. The mutual potential energy of two electric charges varies inversely as their distance and we should therefore expect electric energies in the nucleus to be 100,000 times larger than in the atom. In the latter, they are of the order of a few electron volts, and in the nucleus we should therefore expect between any two particles forces of the order of less than a million volts. But the attractive energy that holds any one particle in the nucleus is, in general, of the order of 6 to 8 million volts. Both these arguments show that inside the nucleus forces of a different kind must be at work and that they act both on protons and on neutrons. We call them "nuclear forces", and to obtain the precise laws of the nuclear forces is one of the central problems of nuclear physics, which is not, as yet, completely solved.

In this respect, the development of nuclear physics is very different from the story of the atom. As soon as the atom was found to contain electrically charged particles there was never really any doubt that the main forces between them are due to their electrostatic interaction and follow Coulomb's law. The difficulty in that case arose from the fact that the general laws of mechanics had to be refined before they could be applied to the atom, as we saw in Chapters 7 and 8. On the other hand, all evidence so far available indicates that the general laws of quantum mechanics, which are so well confirmed in the atomic domain, are still applicable on the smaller scale of the nucleus. But we have no precise knowledge of the forces, and have to deduce all our information about them from observations on the behaviour of nuclei, for example, in their collisions.

In this way, some very general features of these forces have become evident. In the first place, we have already noticed that, to account for the binding energies of nuclei, the forces must be strong, i.e. very much stronger than electric forces between single charges over corresponding distances. One also knows that the forces must be of short range, that is, they diminish with increasing distance very much more rapidly than electrostatic forces. Approximately, this follows already from the old observations of Rutherford that the deflections of alpha-particles passing through matter can be accounted for by the Coulomb force due to the electric charge of the nucleus. In fact, more detailed experiments show that Coulomb's law is valid even when the alpha-particle approaches the nucleus to a distance which brings the two almost into contact. In other words, if the nucleus and the alpha particle are separated by only a very small distance the nuclear force is negligible compared to the electric force. The same conclusion follows from the fact that the potential barrier sketched in Figure 62 predicts very accurately the right lifetime for the emission of alpha-particles if the point *B* is chosen to lie at a radius equal to the combined radius of the alpha-particle and the nucleus.

A more quantitative estimate of the range of the forces, i.e. the distance over which they are capable of acting, is obtained by observing the scattering of neutrons by protons. In the theoretically simplest form of this experiment, a neutron beam passes through hydrogen gas, and one observes how many neutrons are deflected from their straight path and through what angle. If we were dealing with a "long-range" force of the Coulomb type, we should expect that those neutrons which happen to pass at a fair distance from a proton will be deflected only slightly, those which come closer suffering greater deflections. Therefore we would expect many particles to be deflected slightly and a few of them by large angles, as in Rutherford's experiment on the scattering of alpha-rays.

In fact, we observe that very few neutrons happen to hit hydrogen nuclei but that about as many are scattered through large as through small angles. This is precisely the behaviour which one expects for the scattering of waves by a very small localized disturbance, smaller in dimensions than the wavelength. We may conclude that the effect of the nuclear force is limited to a distance rather less than the wavelength of the neutrons used in these experiments, which go up to a few MeV. From the proportion of neutrons which are scattered in passing through a given amount of hydrogen and from the way this proportion varies with the energy of the neutrons one, can in fact, deduce that the forces extend only over a distance of about 2.10^{-13} cm., a small distance even on the nuclear scale.

We can similarly study the passage of protons through hydrogen gas. Here, indeed, we do get many small deflections due to the electric repulsion between the moving proton and the protons in the hydrogen gas, but at high energies and large angles, one obtains additional scattering due to the nuclear force. The interesting result of this experiment is that the nuclear force between two protons seems to be precisely the same as that between a neutron and a proton.

Since a neutron and a proton can be held together by their mutual interaction to form a deuteron, this result would, at first sight, suggest that the analagous problem of the mutual interaction between two protons should also lead to a bound state in which the two protons are tied together. They would then form a nucleus of mass 2 and charge 2, in other words an isotope of helium of atomic weight 2. This, in fact, is not true.

The reason is to be found in the exclusion principle of Pauli, which was discussed in Chapter 8 for electrons, and which is also valid for protons or neutrons. According to this principle, two protons may not be in precisely the same state of motion if their spins are also parallel. On the other hand, neutron and proton are different particles, and thus the exclusion principle does not relate their states to each other. It turns out that the only bound state of the deuteron is one in which the two particles move in the same way and have parallel spins; for two identical particles, e.g. two protons, such a state would violate Pauli's principle.

There is every reason to believe that the nuclear force between two neutrons would also be the same as between two protons or between a neutron and a proton. This cannot be tested directly since it is not, in practice, possible to observe the collision between two neutrons (we never dispose, in practice, of sufficiently many neutrons to put them up as a set of targets into which to fire further neutrons with any hope of a direct hit). However, we know that the neutron-neutron and proton-proton forces must be very similar because each nucleus shows a behaviour very similar to that nucleus which we obtain by exchanging all neutrons for protons and vice versa. For example, the binding energy and the behaviour in collisions of the hydrogen isotope of mass 3 (which is extremely rare in nature but can be made in the laboratory) which consists of one proton and two neutrons, is very similar to the helium isotope of the same mass, containing two protons and one neutron. The latter is somewhat less strongly bound, but this difference amounts exactly to what we would expect from the mutual repulsion of the two protons. This and similar comparisons show that the nuclear forces are perfectly symmetrical between neutrons and protons.

We find therefore the same force between two neutrons, two protons, or one neutron and one proton, and this rule, which is called the "charge independence" of the nuclear forces, leads to a valuable simplification of the problem. More details of the force are disclosed by experiments at higher energies of the order of a few hundred MeV, which have become possible by means of the high-energy machines to which I have referred. At these energies, the wavelength of the particles in the collision becomes less than the range of the forces, and more complicated diffraction patterns result.

One observation is of special interest. At several hundred MeV the kinetic energy of the particles is, in fact, much larger than their mutual potential energy at close approach, which is likely to be about 30 to 50 MeV, and therefore we would expect the incident particle to be deflected only slightly. In other words, when bombarding hydrogen with neutrons of such high energies, we would expect to find most of the scattered neutrons to go nearly in the forward direction, deflected only by small angles.

The experiments show, indeed, that many of the deflected neutrons go almost exactly forward, but an equal number are deflected through large angles, the proton being thrown almost exactly in the forward direction. This is reasonable only if the forces are in part "exchange forces". By this we mean forces which result in the particles exchanging their identity, so that the particle which goes forward has in fact become the proton and carries the electric charge, whereas the particle which was hit and recoils only slightly has become the neutron. The existence of such exchange forces had, in fact, been predicted, long before high-energy experiments were possible, from the trend of the binding energies of heavier nuclei; it would take us too far to review the reasoning that led to this.

There is much detailed information about the forces, which shows them to be of a rather complicated nature. For example, it is known that the nuclear force between two particles depends on whether their spins are parallel or opposite (it has already been mentioned that neutron and proton attract each other sufficiently strongly to form a bound state only when their spins are parallel). When their spins are parallel, the force between them is not, like the electric force between two charges, an attractive force pulling the two particles towards each other, but it acts in a different direction trying to align them so that the line joining them becomes parallel to their spin direction. There are also indications that the force depends on the direction of the angular momentum with which the particles revolve about each other relative to the direction of their spins. There are other complications.

SHELL MODEL. EXCITED STATES

Even if we knew the laws of the nuclear force precisely, it would still be a difficult problem to work out the motion of the neutrons and protons in any but the very simplest nuclei. This difficulty is similar to one that also arises in the case of atoms. There we had found it possible to classify the possible states of an atom by thinking of the electrons as moving independently of each other. Each electron was then supposed to move in a potential field which represented the attraction of the nucleus and the average repulsion due to all the other electrons. Similarly, we may approach the nucleus by thinking of each neutron and proton moving independently of the others in some potential field, which represents the average attraction due to all the other particles together. Again as in the atomic case, the lowest state is one with no rotation, and by Pauli's principle we can accommodate two neutrons (with opposite spins) and two protons. If another particle is added to this, it has to go into the next level and will therefore not be so strongly bound. From this point of view, therefore, a nucleus consisting of two neutrons and two protons, which is just an alpha-particle, should represent a particularly stable group, so that a particularly high energy is required to remove a particle from it, and much less energy is gained by adding a further particle to it. This corresponds with fact, and it shows the reason why alpha-particles occur naturally as decay products from radium and other radioactive nuclei.

One can discuss in a similar way the large nuclei, and this leads to the nuclear "shell model" which has been very helpful in understanding many regularities in the list of nuclei. However, it remains a fairly crude model, far more so than for the atom, and this is not surprising. Because of the short range of the nuclear forces, a neutron moving in the nucleus, is, in practice, not subject to forces exerted by all the other particles but only by a few in its immediate neighbourhood, and therefore its motion is influenced far more strongly than that of an atomic electron by what its neighbours happen to do. This influence is not allowed for in the shell model, which treats the particle as moving independently of the others.

So far we have talked mainly about nuclei in their normal state. But like an atom, a nucleus can also be in an excited state of higher energy. Like the atom, it will in general get rid of this energy in the form of one or more photons, and return to the normal state. Hence, whenever, nuclei are violently disturbed, we expect to find high-energy photons, i.e. gamma-rays.

Nuclei are disturbed by radioactive decay, such as the emission of an alpha-particle, and this is the origin of the gamma-rays which were found to accompany radioactive decay in some cases. They may also be disturbed by external bombardment, and this gives us the possibility to study, in the case of almost all nuclei, the emission of gamma rays, and hence the energies of their excited states.

Other evidence about these excited states comes from the energy loss of particles in nuclear collisions (the energy lost by the particle has gone to excite the nucleus) and from the energy balance of reactions in which one or more of the nuclei end up in excited states.

The study of nuclear energy levels or "nuclear spectroscopy" has thus led to an enormous collection of data, which may be used to throw further light on problems of nuclear forces and nuclear motion in detail.

BETA RAYS

We have not so far mentioned the "beta rays" which have also been well known since the earliest work on radioactivity. By this name one refers to the rays which are given off in some nuclei and which Rutherford showed to consist of fast electrons. Where do these come from? One might think that they were also part of the structure of the nucleus, but this would then contradict the conclusion that nuclei consist only of neutrons and protons, which has been amply confirmed in many ways. Besides, it is very difficult to believe that electrons could be confined to a region as small as the atomic nucleus since, by the uncertainty principle, they would then have to have kinetic energies equal to many hundreds of MeV, and if there existed attractive forces strong enough to confine the electrons in spite of this, they would have shown up in many experiments.

One must therefore conclude that these electrons are produced when the beta-decay of a nucleus takes place and that their negative charge is balanced by a neutron turning into a proton. This is not however the whole story. Beta-decay usually takes place from a nucleus which has been undisturbed for a long time and which we must therefore assume to be in its state of lowest energy. Its energy is therefore well defined. Its beta-decay can take place only if the final nucleus, which has one proton more and one neutron fewer than the original one, has lower energy. This nucleus will also end up in a well-defined state, and the energy difference E between these two states is then available to the electron. Allowing for the rest energy mc^2, the remainder, $E - mc^2$, should appear as its kinetic energy. In other words, all the electrons resulting from the beta decay of a given nucleus should have the same energy.

But this is not found. If we measure the energies of the electrons coming from a given radioactive substance, they range, instead, all the way from nothing to a maximum, which, whenever the energy difference E can be determined directly from other reactions, is equal to the expected available energy $E - mc^2$. In each beta-decay process, therefore, a variable amount of energy seems to disappear.

A similar contradiction is found if one considers the total angular momentum, which should not change in such a process. Suppose for the sake of argument that the nucleus contains an even number of neutrons and protons together. Each of these, as we know, has a spin of half a unit. An even number of such particles therefore add up to a whole number of units of angular momentum. Their motion about each other also gives a whole number of units, as we saw in Chapter 8, and therefore, before the beta-decay, we had altogether a whole number of units of angular momentum. The same is true of the nucleus remaining after the decay. But in addition, we have obtained an electron, which itself has a spin of half a unit. In other words, half a unit is missing from the balance, and this cannot be compensated for by any motion of the particles, which will always contribute a whole number.

In the face of these discrepancies, Pauli suggested the production, together with the electron, of a new particle which escapes detection. This must again be neutral (if it carried an electric charge it would easily show up in any of the usual detectors); it must also have a spin of half a unit. This assumption would immediately restore the balance of angular momentum, and it is perfectly reasonable to expect that in a process in which two particles are produced the available energy will be shared between them in a manner that varies from case to case. This new and very elusive particle was given the name "neutrino".

Detailed consideration of the way the energy is shared between the electron and the neutrino showed that its mass must be much less than that of an electron. It is probably zero, like that of a light quantum. It seemed at first somewhat unsatisfactory to invent such a new particle, which had never been seen directly, merely, as it were, to balance our books. It acquired a little more reality when careful experiments about the recoil of the decaying nucleus in a beta-decay process showed that also the momentum of a nucleus did not balance exactly the momentum of the electron that was produced, and that the lack of balance again agreed with the hypothesis that some particle had been thrown out with a momentum just equal to that of a particle travelling with the missing energy.

Nevertheless we shall never be absolutely sure of the existence of the neutrino until we have succeeded in observing a process caused by this particle. In principle, this is obviously possible. If it is possible for a neutrino to be emitted in a nuclear process, the converse must also be possible, namely that a neutrino colliding with a nucleus would be absorbed, either together with one of the atomic electrons which are in the neighbourhood, or leading to the production of a new electron. But it is also known that such events would be exceedingly rare because the beta-decay itself is an exceedingly slow process. The mean life of a beta-decaying nucleus may be as long as hours or weeks or more according to the available energy, and this is an extremely long time compared, for example, to the time a neutrino would take to travel past a nucleus, the latter time being of the order of 10^{-23} seconds. Accordingly, in the inverse process, a neutrino would have to pass close to very many nuclei indeed before it had a reasonable chance of once causing an inverse beta process. It has, for example, been estimated that a neutrino would be able to pass many times right across the interior of the earth before it would once make a collision.

In spite of these tremendous odds, an indication of the inverse beta process has now been found. It will, however, take some time before this discovery is established beyond doubt.

Beta-decay is not confined to the naturally radioactive elements but it occurs whenever we produce, by bombardment or in other ways, a nucleus which has relatively too many neutrons in relation to the number of protons, so that its energy could be reduced by more than mc^2 if we change a neutron into a proton. Even the neutron itself is radioactive, since, as has been mentioned before, it is slightly heavier than a proton plus an electron. It decays into a proton, an electron and a neutrino with a mean life of about 20 minutes.

Conversely, if a nucleus contains too many protons, a proton can change into a neutron, with the emission of a positive electron and a neutrino. This process was discovered by the Joliot-Curies and it was the first example of artificial radioactivity, i.e. of making by artificial bombardment new unstable nuclei which emitted beta-rays.

The existence of both negative and positive beta-decay explains why each element in nature has only a few stable isotopes. If we think of the neutrons and protons in a nucleus as filling separate energy levels or "shells", then if we have too many neutrons in a nucleus the Pauli principle will require us to put the last ones into rather highly excited states or high shells, whereas the fewer protons will fill only the lowest levels. In that case, we can obviously gain energy by transforming a neutron into a proton, since this would allow us to drop it from the highest occupied neutron shell to the lowest level not completely filled with protons. In other words, energy can be released by changing a neutron into a proton and that means beta-decay becomes possible. Conversely, if there are more protons than neutrons, we could liberate energy by changing a proton into a neutron by positive beta-decay.

In fact, if one takes this argument literally, one might expect that nuclei will be stable only if they contained approximately equal numbers of neutrons and protons. For light elements this indeed is true. For example, the normal form of oxygen, as we have seen, contains 8 neutrons and 8 protons. But for heavy nuclei the number of neutrons generally exceeds that of protons. In the most stable form of lead, for example, there are 82 protons and 126 neutrons. The reason for this lies in the electric repulsion of the protons which give to the nucleus a high electric potential energy which rises with the number of protons. This shifts the balance in the direction of having fewer protons.

The actual neutron-proton ratio represents a compromise between the tendency to equalize the numbers so as to accommodate as many particles as possible in the lowest shells and the tendency to reduce the proton number in order to reduce the electrostatic energy. The actual trend of stable nuclei can be explained very satisfactorily from this reasoning.

FISSION. SOURCES OF ENERGY

As we have just seen, the heavier nuclei have a considerable electrostatic energy due to the electric charge of the protons which they contain, and we can therefore gain energy also by breaking such a nucleus in two, thus reducing the charge on each part. Each part so produced will be beta-active and increases its charge somewhat to make the numbers of protons and neutrons more nearly equal. For most nuclei this is only a hypothetical experiment because to split such a nucleus in two is an operation which cannot be carried out without first investing a lot of energy even though, in the end, we shall get back even more than we have put in. The easiest way to carry it out would be as illustrated in Figure 65.

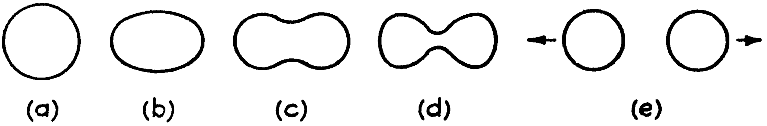


Fig. 65. Fission.

The nucleus is normally nearly a sphere as in 65*a*. If we pull it out a little into an oval shape, as in 65*b*, we have to do work against the attraction of the neutrons and protons on each other, which tries to keep them in the most compact shape, i.e. a sphere. At first, we gain only a little energy from increasing the distance between the electric charges. But if we have gone far enough, the further increase in distance will more than outweigh the loss of compactness and once we have achieved a shape like *c*, the repulsion between the ends will push the nucleus into a shape like *d* and eventually the two parts will separate. They will then be thrown apart with considerable force owing to the electric repulsion of the fragments.

This process is known as "fission". If the charge is high enough, greater than, perhaps, 120 proton charges, the whole process starting from *a* will go with decreasing energy, in other words it will take place immediately and without any external intervention. Such nuclei could never exist for long enough even to be observed.

Even for smaller charge, there is a slight chance of this fission process taking place spontaneously. This is due to the phenomenon of the passage through a potential barrier, which we have discussed at the end of Chapter 8. In quantum mechanics a nucleus can pass through a state of high energy by "borrowing" the energy required for this, provided the loan can be repaid after a short time and there exists therefore a small chance of the nucleus borrowing enough energy to get through stages *a* to *d* and breaking up. This process of "spontaneous fission" is measurable, even in uranium, the heaviest element occurring in nature. There it is rather a rare process, much less frequent than the rather slow alpha-decay, so that most uranium nuclei will have emitted an alpha-particle before they undergo fission. Spontaneous fission is one of the reasons why heavier elements do not survive in nature.

The fission process is assisted by neutron bombardment. If even a slow neutron hits a uranium nucleus, it is attracted by the other particles and therefore enters the nucleus with considerable kinetic energy. This energy will be shared with all the other particles and will result in a violent and irregular motion of the whole nucleus. In the course of this motion, the nucleus may well be distorted into shapes like Figure 65*b* or *c* and in that case fission will result.

These vibrations of the nucleus continue however only for a short time because any vibrating system containing charged particles emits electromagnetic radiation, just like an excited atom and thereby loses energy. The question is therefore whether in the short time for which the vibrations last, there is a chance of fission. This depends both on the violence of the vibration (and this in turn depends on how strongly the extra neutron is attracted) and also how strong is the tendency to fission. In this respect the lighter isotope of uranium, of 235 units of mass, is more suited to fission than the heavier and more abundant isotope of mass 238. The former has a smaller size for the same charge, and its tendency to undergo fission therefore is stronger. At the same time its attraction on an extra neutron is greater and therefore by capturing a neutron it is set into more violent vibrations. As a result, if a nucleus of uranium 235 captures a slow neutron, it will nearly always undergo fission. In the case of a nucleus of uranium 238, this will hardly ever happen. Uranium 238 can be split if we use fast neutrons which increases the violence of the vibrations. In several other elements fission can be produced by sufficiently fast neutrons.

When the two fragments have separated as in Figure 65*e*, they will, as a result of being torn apart, again be in violent internal motion and this is equivalent to the heat motion of a drop of liquid as discussed in Chapter 5. Just as the heat motion of the molecules in a liquid causes it to evaporate, i.e. to throw out some molecules, the motion of the fission fragments may cause them to throw out one or more neutrons. The importance of this is that these neutrons may collide with other uranium nuclei in turn, causing them to split and to produce more neutrons. In this way, the number of neutrons rises rapidly. If we assume, for simplicity, that each fission produces two neutrons and that the loss of neutrons due to causes other than collisions with uranium nuclei is unimportant, then the first fission would produce two neutrons. Each of these would lead to another fission and thus to two more neutrons, so that the second "generation" would consist of four, the next of eight, the next of sixteen neutrons and so on. In this way, the total number of neutrons and therefore also the total energy released would rapidly rise to catastrophic proportions unless steps were taken to keep the process under control.

The importance of these facts for the possibility of atomic bombs and for reactors producing atomic power in a controlled manner is evident. If we had here been concerned with the importance of nuclear physics for practical purposes, we should have regarded the whole of this chapter merely as a preparation for describing nuclear fission, but since our purpose is to learn about the laws of nature we must regard the discovery of fission as only one small episode in a long and continuing story.

The discovery of atomic energy has, however, also made important contributions to the advance of our knowledge of the nucleus, not only because it led to an increased interest in nuclear physics. Nuclear reactors in which, as we have seen, very large numbers of neutrons are being liberated, and which are therefore extremely powerful sources of neutrons, have to be added to the list of machines available for producing intense beams of projectiles with which to bombard atomic nuclei and study their behaviour.

The energy released in the fission process, and utilised in atomic reactors and atomic bombs, comes from the breakdown of nuclei found in nature which are too large for economy, so that energy is gained by breaking them up.

At the other end of the scale, energy can be gained by combining nuclei which are too light. This source of energy is also of the greatest practical importance to us since, in the last resort, it is the source of the heat of the sun (and of the stars) without which life would be impossible and from which also all other sources of power available to us are directly or indirectly derived. In the interior of the sun collisions between atomic nuclei take place which result in the combination of lighter atoms into heavier ones, in particular the formation of helium and other elements from hydrogen. What is required for this process?

In the breakdown of a heavy nucleus such as uranium, we are concerned with one that is, in itself, explosive and requires only the slight disturbance caused by a slow neutron as a "trigger" action. In the combination of two light nuclei, it is necessary to bring them close together and this is opposed by their electric repulsion. It has already been remarked that if we try to assist this process by using protons or other charged particles which we have accelerated in the laboratory, most of them will have been slowed down by friction before they have had a chance of a close collision and once they are slow further collisions are prevented by the electric repulsion. In this slowing-down, their energy is converted into heat. The only chance of converting a substantial amount of matter by such collisions is to raise the temperature to such a point that the irregular heat motion of the nuclei gives them a velocity high enough to collide in spite of the repulsive forces. At such a high temperature the nuclei do not lose their energy by friction, but as they pass through matter they tend to gain as much energy as they lose, maintaining on the average a kinetic energy of $\frac{3}{2}kT$ (k is Boltzmann's constant and T the temperature), as we had seen in Chapter 4.

In the interior of the sun, the temperature is about 20 million degrees Centigrade and, at such a temperature, the kinetic energy of any nucleus is 3,000 electron volts. This equals the potential energy of two protons or deuterons at a distance of 5.10^{-11} cm. Therefore the protons can easily approach to within this distance in spite of the forces between them. This is not yet close enough for a direct collision, since the radius of the heaviest nucleus is only 10^{-12} cm. and lighter nuclei are still about five times smaller. However, two facts help: firstly, the possibility, discussed at the end of Chapter 8, of the penetration of particles through a potential barrier which we had also recognized as important in the process of alpha decay. Secondly, in the irregular heat motion the atoms do not all have the same velocity, but some are faster and others slower. There will therefore always be a fraction of the atom which happen to have enough speed for a close collision.

But even in the extreme conditions of the interior of the sun, such collisions are very rare. Only perhaps one atom in a million will make such a collision during one year, otherwise the sun would already have "burnt out" before now.

If one wants to produce such "thermo-nuclear" processes artificially, one must therefore produce temperatures substantially higher than those of the sun in order to get substantial returns in a reasonable time. This rules out, as far as one can see, any possibility of using this process for a steady and controlled release of energy, as is possible with the energy from fission, since no materials exist which could withstand these extreme temperatures for even a few seconds. It is different with an explosive release of energy in which the walls of the vessel containing the ingredients and the rest of the equipment will, in any case, be destroyed. While this book is being written, it has been disclosed that both the United States and Russia have succeeded in setting off explosions of such "thermonuclear bombs" While the details have not been revealed, the principle is presumably that the energy released by a fission bomb is used to heat a suitable mixture of light elements to a temperature similar to, or higher than, that at the centre of the sun and thereby starts a thermo-nuclear reaction.

In exploring the structure of matter in our search for the basic laws, we have found it to consist of atoms. We found that the nature of the atom depended on its nucleus, whose mass was essentially that of the atom and whose electric charge determined the number of electrons it normally carried and through this, its chemical behaviour, as well as its ordinary physical characteristics. We have seen that nuclei were built of neutrons and protons in known proportions, and we acquired some knowledge of the behaviour of neutrons and protons in the nucleus and of the forces between them, though this knowledge is, as yet, very far from being complete.

We have realized the old dream of the alchemists to change one chemical element into another, at least in small amounts, though the value of any precious material so produced is far overshadowed by the power we have learnt to release and which, according to our own wisdom, we may use for our own good or for destroying each other.

Two kinds of questions still remain unanswered. The first of these is about the origin of matter. How were the nuclei formed which constitute the matter around us? Why are some elements more common than others? Why are some isotopes of each element, much more abundant than others?

These questions are connected with the origin of our planet and the history of the universe. At the present time, no changes in the atomic composition of the matter around us take place apart from the small surviving amounts of unstable elements like radium and uranium. But presumably, in the distant past, the earth and the other planets were, together with the sun, a mass of very hot matter, in which by thermo-nuclear reactions larger nuclei could have been formed from small amounts. Speculations of what came before this bring us into the field of cosmology, the science of the structure and origin of the universe. There is no place in this book for a review of the main problems of cosmology and, in any event, they cannot yet be regarded as part of the firmly established laws of nature. The field of cosmology abounds with fascinating speculations, but few conclusions are, as yet, free from controversy.

Another kind of unsolved problem concerns the precise nature of nuclear forces and the question of what other particles might exist besides those we have already met, i.e. the electron, proton, neutron, photon, neutrino and their combinations. The facts to be described in the last chapter, which will bring us up to the front line in the search for the basic laws, will also have a bearing on these questions.

11

Mesons and Other New Particles

COSMIC RAYS AND TOOLS FOR THEIR STUDY

THIS chapter is concerned with a number of new particles, most of them discovered only in the last few years. The laws which determine their behaviour are, as yet, known very incompletely, but it is already clear that there are important connections between them and the laws of the nucleus. It is, in fact, almost certain that we must expect to find in a more complete study of the new particles important clues to the remaining problems about nuclear forces.

Most of them were first discovered amongst the products of the cosmic radiation. It will therefore be useful to begin the story with a brief account of this phenomenon.

As has been discussed before, whenever charged particles pass through a gas they collide with the gas atoms and knock off electrons. This leaves free electrons and positive ions, which can conduct electricity. The most primitive way of demonstrating this is to charge an electroscope of the kind sketched in Figure 8 and leave it in a position where the outer plate, *P*, is close to, but not in contact with, another metal plate connected to the earth. If then radium or some other source of charged particles is placed nearby, the electroscope gradually loses its charge. This is made apparent by the leaves gradually dropping to the vertical position. The reason is that electric charge has been conducted away to the earthed plate by the ions in the air. This was, in the early days, a convenient method for measuring the amount of radiation present. It is still used in almost this form in instruments which record the amount of radiation in a laboratory to keep a check on the total radiation to which personnel may have been exposed.

In order to make this method more sensitive, one tried to improve the accuracy of the electroscope and also to improve its insulation, so that the picture was not confused by current carried through the wall of the electroscope. It was then found, however, that there still appeared to be a small leak of charge even in the complete absence of any radioactive material. The electroscope was still losing its charge at a slow but steady rate. In the course of studying this phenomenon, Hess discovered that this background ionization could be reduced if the electroscope was surrounded by heavy walls, or placed underground or under water. On the other hand, it becomes much more intense, i.e. the electroscope loses its charge more quickly, if it is taken to a high altitude, either on a mountain or in an airplane or balloon.

He concluded that what he had found must represent a radiation which comes from above and, since one can hardly imagine a source for it in the upper atmosphere, it must come from outer space and be capable of passing right through the whole atmosphere. These deductions were completely confirmed by later work. As the detectors for charged particles have improved in sensitivity and in their ability to distinguish different kinds of particles, our knowledge of the nature of the cosmic radiation has improved, though we still do not know with certainty where it comes from.

At first, it was taken for granted that cosmic rays consist of electromagnetic radiation, i.e. are in effect gamma-rays, which in their passage through matter can liberate electrons by Compton effect or photo-effect. Charged particles, on passing through air lose energy by friction. To reach sea level at all or to get through concrete walls or lead plates through which cosmic rays were known to pass charged particles would have to have energies which, at that time, seemed too fantastic to be taken seriously. However, we know now that gamma-rays cause only a small part of the ionization which we observe at sea level and most of it is due to charged particles.

Their existence can be demonstrated in many ways. A typical method is to use counters of the type described in Chapter 10, which are essentially tubes filled with a suitable gas and containing a wire with an electric potential difference between wire and tube, such that the small amount of ionization caused by a passing particle is enough to start a small electric discharge. The discharge of such a counter shows merely that a particle has passed through it, but does not tell us much about the nature of the particle. More information is obtainable by using counters "in coincidence". This means that two counters like *A* and *B* in Figure 66 are placed close to each other, and the recording device is arranged so that it only records events in which both counters are discharged at precisely the same time. When this happens either a single particle has passed through both or different particles have each passed through one counter. The second alternative which gives "chance coincidences" is usually

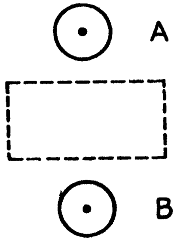


Fig. 66. Counters.

uninteresting. If the times are determined accurately and if the rate at which particles are passing is kept low, the number of such chance coincidences can be kept small. In any event, it can be allowed for.

In this way one can study whether or not particles are capable of passing through, let us say, a lead block of a given thickness. For this purpose, one merely has to place the lead block between the counters as indicated in the figure and observe how much the number of true coincidences goes down. More refined techniques specify the events in greater detail. For example, one may place three counters in a straight line with blocks of various substances between the first and second and between the second and third. If one then records the number of events in which the first and second but not the third counter has operated, this shows how many particles have passed through the first block but stopped in the second.

Another important detecting device is the cloud chambers also already mentioned in Chapter 10, in which one can photograph the tracks of passing particles. In general, the rate at which cosmic rays arrive is so rare that one has to expose a cloud chamber of reasonable size many times, at least at sea level, to have a good chance of seeing at least one track. This disadvantage of having to take many useless photographs for each useful one can be avoided by combining the counters and the cloud chamber. See Figure 67. Counters *A* and *B* are placed in such a way that a particle which has passed in a straight line through both counters must also have passed through the cloud chamber *C*, and the coincidence between the discharge of both counters is used to "trigger" the cloud chamber mechanism. This means that the cloud chamber operates, i.e. the water drops are condensed on the track, only after the particle has gone, but this does no harm, since the ions which the particle has produced do not disappear very quickly and there is time enough to produce and photograph the "vapour trail". In this way it is possible to ensure that each photograph will show at least one track of a particle capable of penetrating through both walls of the cloud chamber. Moreover, this technique also gives a great advantage in allowing us to select particular types of particle, or particles to which certain selected events have happened. For example, one may

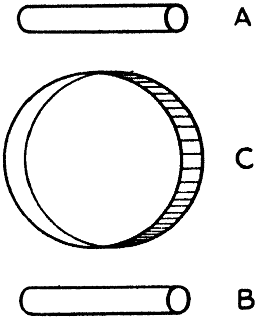


Fig. 67. Counter-controlled cloud chamber.

place another block of lead between the chamber and counter *B*, or below *B* followed by a third counter, and in this way select particles which have still enough energy to pass through the block of lead, or conversely particles which are known to stop in it.

More information can be obtained by placing the cloud chamber between the poles of a strong magnet so that each track becomes curved, and from its curvature it is then possible to obtain information about the momentum of the particle.

A third important technique uses photographic plates. It has been known for a long time that a charged particle, for instance an alpha-particle, passing through the sensitive layer or "emulsion" of a photographic plate, which is later developed, leaves a row of black dots. This is again due to the ionization of some atoms inside the small grains of silver salt contained in the photographic emulsion. The electrons thereby released in such a grain cause chemical changes which make the grain "developable", i.e. convertible into silver in the process of development. This is precisely like the normal action of the photographic plate when exposed to light, except that it is then the photo-electric effect of the light which releases electrons, rather than the collisions of the passing alpha-particle with atoms.

Under a microscope, the track of a fast particle is therefore visible as a row of dots on an ordinary photographic plate. However, if the speed of the particle is increased, it produces fewer ions on its way, because the electric disturbance it produces in passing through an atom is over so quickly that there is a greater chance of the atom remaining undisturbed. Hence fast particles make very few marks on a normal photographic plate and their tracks are hardly visible. However, there are now special plates which do give visible tracks even from particles travelling practically with the velocity of light.

In its application to cosmic rays this method has two obvious advantages. One is that the plates can be exposed for long periods and will thus record all the tracks that have passed through the emulsion in that time. Therefore the low intensity, i.e. the small number of particles passing through the plate per minute, is no disadvantage. During the exposure the plate needs no attention. The other advantage is the small weight, which makes it easy to carry a large number of such plates in a balloon or in a rocket to a great altitude where the cosmic rays are much more numerous and where there are also types of radiation present which cannot be observed at sea level. A further specific advantage will become evident later.

KNOWN PARTICLES

A closer study of cosmic rays with all these techniques has shown up great complexity. Some of the tracks which were found turned out to be particles of familiar kinds, in particular electrons and positrons. Together with gamma-rays (i.e. high energy photons) they make up what is known as the "soft component" of cosmic rays. This means that they are stopped by comparatively small thicknesses of heavy material such as lead plates. We know already that any charged particle, in particular an electron, in passing through matter loses energy by friction as it passes through atoms. But if that were all that could happen to it, the penetrating power might still be very large for an electron of very high energy, since it has more energy to lose. In addition, however, electrons tend to lose energy by radiation. When an electron happens to pass close to the nucleus of an atom it is deflected suddenly by the strong electric force and, in the course of this deflection, it tends to lose some of its energy in the form of gamma-radiation. The photons produced thereby may, in passing through atoms produce pairs of electrons and positrons as we saw in Chapter 9. The net result is therefore that part of the energy of the original electron has been converted into one or more pairs of correspondingly lower energy. These, in addition to being slowed down by friction, may in turn radiate and produce more photons and more pairs. This process is known as a "cascade". As the number of particles in the cascade increases, the energy of each of them becomes smaller, so that the growth of the cascade ceases when the electrons and positrons have such low energy that they are stopped by friction before they have a chance to radiate and when the photons have such low energy that they are no longer capable of producing pairs.

In addition to this "soft component", cosmic rays also contain protons. These also are being slowed down by friction, but they do not produce gamma-rays to any appreciable extent since they are much heavier and are therefore deflected much less violently by the electric forces they encounter in passing through an atom. On the other hand, a proton does suffer a violent disturbance if it makes a head-on collision with an atomic nucleus.

In passing through air a particle travels, on the average, a few hundred metres before it happens to score a bull's-eye, i.e. collides directly with the nucleus. If this happens to a proton, the impact shatters the nucleus into fragments and the proton loses an appreciable part of its energy. Such collisions of protons with nuclei also give rise to various new kinds of radiations about which more will be said later.

Protons, in fact, make up the bulk of the cosmic rays which reach the earth from outer space. That this "primary" radiation consists of charged particles can be shown by comparing the cosmic rays at different places on the earth. There is much more radiation observed near the magnetic poles than near the equator, and this is because in travelling through space towards the earth these protons, like any charged particles, are deflected by the magnetic field of the earth and, unless they have very high energy, fail to reach the earth. Slow particles can reach the earth only if their line of approach happens to coincide with a straight magnetic line of force. In that case they are not deflected since the force of a magnetic field on a current depends on the part of the magnetic field at right angles to this current. Now if we compare the earth to a little magnet, as in Figure 13a, we recognize that the only straight lines of force are those which start from either pole and go off in the continuation of the line joining the poles. Therefore the only chance of the slow proton reaching the earth is at the magnetic poles. In any other place there is a minimum energy which a proton must have to be able to arrive there in spite of the magnetic deflection. This minimum energy is greater the further away we are from the poles. Hence the cosmic-ray physicists can use the earth as a gigantic instrument for separating the effects of primary protons of different energies. If one is interested in low-energy protons, one must therefore make one's observations close to the magnetic pole. To isolate the effect of protons of extremely high energy one works near the magnetic equator.

It is now known that, besides protons, the radiation striking the earth's surface from outside also contains other nuclei, ranging from helium at least to nitrogen and probably further. In fact it would seem that the proportions in which protons, i.e. hydrogen, and other elements are mixed in the cosmic radiation are quite similar to the proportions in which they are found in the atmosphere of the sun and in other parts of the universe. This is one of the important clues which we have about the origin of the cosmic rays, but no complete theory is, as yet, available.

These heavier nuclei are found only at very great altitudes because when such a high-speed nucleus collides for the first time with an atomic nucleus in the air it is likely to be shattered into separate neutrons and protons which continue on their way separately, together with any fragments which they have knocked off the nucleus.

As a result of such collisions, neutrons are also contained in the cosmic radiation as it proceeds downwards through the atmosphere.

We know that neutrons cannot form part of the rays arriving from outer space because of their limited lifetime, which was mentioned in Chapter 10.

MESONS

In addition to these familiar objects, cosmic rays also have shown up new particles which were not previously known. The first evidence of these newcomers was obtained in the early 1930's when it was proved that there were charged particles which could traverse thick lead plates without apparently being deflected from their straight path or producing a cascade or any other kind of additional track. They could not be electrons, since any electron of sufficient energy to pass through such a lead plate ought to produce a cascade. They could not be protons because some of them were found to be negative (and as we now know they do not produce nuclear fragments when the lead plate is thick enough to make it certain that some of them must have passed through nuclei).

Later measurements showed these particles to be lighter than protons but much heavier than electrons. How does one weigh such a particle? There are several ways of finding out something about a particle. One is to observe its range, i.e. the distance in which friction brings it to rest. This is essentially a measure of their energy, since the amount of energy lost by friction in travelling a given distance is very similar for different kinds of particles. (This statement is an over-simplification, since the amount of energy lost depends on the velocity, but this variation can be allowed for.) One can also observe the deflection in a strong magnetic field, either in a cloud chamber, or using counters, and this deflection depends on the momentum. One may observe the number of ions which the particle produces along its track, by looking either at the number of waterdrops along the track in a cloud chamber or at the number of dots making up the track in a photographic plate. This tells us something about the velocity of the particle, since its chance of throwing an electron out of an atom depends on the time it spends passing through the atom. Lastly, one can observe the deviations of the track from a straight line since, in passing through each atom, the particle is deflected slightly and therefore as a result of many such small deflections in different directions, the path becomes slightly irregular. This irregular "scattering" depends essentially on the product of the momentum and the velocity of the particle.

In principle, any two such observations, for example, those of energy and momentum, are sufficient to fix the mass of the particle.

But since the accuracy possible in practice for these different observations depends on the circumstances, a great deal of ingenuity has to be used to tie down the mass as accurately as desired.

The new particles were found to have a mass about 210 times that of the electron, in other words, a little over one tenth of the proton mass. Such particles are now known as "mesons", from a Greek term which implies their intermediate position between proton and electron.

These new mesons, discovered by C. Anderson and, independently, by Blackett, were found to live for only a very short time. After about two microseconds, i.e. 2 millionths of a second, on the average, such a particle decays, the only visible product being a fast electron. This process is reminiscent of the process of beta-decay. If in this process energy and momentum are conserved, at least one other particle must be produced, as may be seen by considering the meson at rest, i.e. with no momentum. Since the electron comes out with considerable speed, i.e. considerable momentum, conservation requires that an opposite and equal momentum is carried away. In fact, if only one other invisible particle were produced, it would always have to have the same momentum as the electron and therefore the available energy, which is the rest energy of the meson, would always be shared between the electron and the other particle in the same proportions. For example, if, as is in fact true, the available energy is much greater than the rest energy of both the electron and the other particle, both travel almost exactly with the velocity of light and the energy of each is its momentum times the light velocity. In this case, for equal momentum they must have equal energy and each will have exactly half the available energy, i.e. one half of the mass of the meson times c^2.

However, experiment shows that this is not so. As mesons disintegrate, the electron sometimes comes out with more and sometimes with less energy. This can happen only if, besides the electron, at least two invisible particles are produced. The most likely explanation at the present is that the break-up of the meson is accompanied by two neutrinos.

This conclusion makes the process more similar to that of beta-decay. Just as in the ordinary beta-decay, the neutron is changed into three particles, namely a proton, an electron and a neutrino, so here the meson is changed into an electron and two neutrinos.

This similarity also extends to the lifetime. The lifetime of a nucleus which can undergo beta-decay depends, in a known way, on the energy released in the beta process, and if one works out what would be the lifetime of a state in which the whole rest energy of a meson, i.e. about 100 MeV, was released, one finds the correct result of about two microseconds.

Such mesons make up a substantial proportion of the cosmic rays at sea level. Many more of them must have been present at high altitudes and have decayed on the way. The electrons resulting from this decay are the main source of the soft component of the radiation, which was discussed before.

It is evident that because of this short lifetime, the mesons must be produced in collision processes in the atmosphere. They cannot be contained in the primary radiation which comes in from outer space, since they would then, with certainty, have decayed during the journey. In fact, one might conclude, at first sight, that even mesons produced in collisions near the top of the atmosphere could never reach sea level, since, travelling even with light velocity they would cover in two microseconds a distance of only 600 metres, which is far less than the height of the atmosphere.

This argument is, however, incorrect and the reason is interesting: many of the mesons have very high kinetic energies so that they travel with a velocity extremely close to that of light. An observer travelling with the same speed as the particles, for whom they would therefore appear to be at rest, should still find their lifetime to be two microseconds. But if the findings of such an observer are translated to those of a stationary observer, by the laws of the Lorentz transformation (see Chapter 6), we have to allow for the time dilation, and to us therefore the particle will appear to live much longer, the increase being in the same proportion as the total energy of the particles to their rest energy. For example, mesons with an energy of 100,000 MeV, which are still common, will, on the average, survive for 2,000 instead of 2 microseconds and in this time they will cover a distance of 600 kilometres.

We see here that the result of the time dilation can be very tangible, not restricted to academic problems like those of observers comparing their stop watches while travelling at speeds close to that of light. Incidentally, it might be worth noting that from the point of view of an observer travelling with the mesons, the same situation would appear to be due to the Lorentz contraction of distances; while such an observer would record the mean lifetime of the mesons to be 2 microseconds, the height of the atmosphere which, to him, rushes past at high speed, is reduced by the Lorentz contraction again in the ratio of 1 to 1,000 in the numerical example, and therefore the meson's lifetime is sufficient to cover that distance.

It might appear curious that the only way these particles can decay is through such a complicated process which results in three particles, but this is more natural if we assume, as seems reasonable, that the meson, like the electron, has a spin of half a unit. It cannot transform into a single electron since energy and momentum could not then be balanced. It cannot decay into an electron and one neutrino since two such particles together would always have an angular momentum of a whole number of units. Therefore the simplest possible decay is that into three particles.

YUKAWA THEORY. MORE MESONS

The discovery of these mesons appeared to give strong support to a suggestion which had been made earlier by the Japanese physicist, Yukawa, on the basis of theoretical speculations about the nature of the nuclear forces. It is therefore appropriate, at this point, to review Yukawa's argument.

Yukawa was trying to find a more satisfactory description of the force between two nuclear particles, for example, a neutron and a proton. This force is known to be a short-range force, i.e. it is very weak if the distance between neutron and proton exceeds about 2.10^{-13} cm. but it does act over distances of this order of magnitude. Now one cannot admit the existence of forces directly linking particles at a distance, however small, because the principle of relativity, which we have found to be a very powerful guide in fundamental physics, does not admit the possibility of any actions transmitted instantaneously, or with any speed greater than that of light. Hence, one must assume that the nuclear force also is transmitted from the proton to the neutron or vice versa through some new agency which, for these nuclear forces, plays a similar part to that played by the electro-magnetic field in the transmission of electric forces. Just as we had seen in Chapter 2 that each electric charge causes a disturbance in the electro-magnetic field, which is transmitted from point to point until it finally reaches another charge, there must, similarly, exist a field which transmits nuclear forces.

However, the fact that nuclear forces extend only over short distances means that the laws of this field must be rather different in detail from those of the electromagnetic field.

Following this chain of reasoning, Yukawa came to the conclusion that such a field must also be associated with particles, which, however, were not particles of zero rest mass like photons, but had to have a mass a few hundred times larger than that of the electron.

Without entering into mathematical refinements, this argument may be understood if we remember the reasoning of Chapter 9, in which we had connected the electric force between two particles with the emission of virtual or "borrowed" photons. One charge may emit a photon on credit and this may be absorbed by another charge nearby. If the distance is small, either long or short-wave photons may be exchanged in this manner, since the loan can be repaid immediately, but if the distance is larger, it will be longer before the loan is repaid and, by the uncertainty principle, only a smaller loan is then possible. This means that only photons of a correspondingly smaller frequency or longer wavelength may be included. It is possible to derive the inverse square law from this reasoning.

Now consider, in comparison, a process in which not photons, but new particles of appreciable rest mass are produced. Then, however long the wavelength, or however small the momentum of these particles, their energy is at least mc^2, where m is their rest mass. Hence, the amount of energy to be borrowed is never less than mc^2 and, by the uncertainty principle, the process can occur only if the loan can be repaid in a time $\frac{h}{mc^2}$. Since the action can be transmitted at most with light velocity, it can cover a distance of, at most, $\frac{h}{mc}$. The production and absorption of such particles will therefore give rise to forces acting over distances of a few times 10^{-13} cm. if the mass is assumed to be a few hundred times the electron mass.

The discovery of mesons in the cosmic rays seemed therefore to confirm these speculations of Yukawa's and, for about ten years, one tried to identify the mesons normally found in cosmic rays with those occurring in Yukawa's theory. However, this view led to a number of contradictions. The most serious one of these was that mesons seemed able to pass right through atomic nuclei without anything drastic happening to them whereas, if Yukawa's theory was right, they ought to be absorbed very readily by nuclei, their mass being converted into energy in the process.

These difficulties were resolved only when Powell and his collaborators discovered the existence of two different kinds of mesons which he called π ("Pi") and μ ("Mu") mesons. The μ mesons were the particles which had been studied extensively before. The π mesons were new. They interacted very strongly with nuclei and altogether behaved in accord with the ideas of Yukawa.

Powell's work was done by means of photographic plates. In studying the tracks of cosmic ray particles in such plates, he found, on occasions, tracks of particles which were slowing down and ultimately coming to a stop. This can be seen by their track becoming blacker and blacker towards the end (a slow particle produces more ions and therefore blackens more grains in the photographic plate, as we have seen) and, at the same time, becoming more irregular (a slow particle is more easily deflected in collisions). But in some cases from the end point of such a track, a new fast particle would start, evidently with considerable energy. Both particles were moving with velocities much less than that of light, as could be seen from the number of grains along their track, and therefore could not be electrons, since they covered much greater distances than would be possible for an electron of corresponding speed. The energy of the second particle was much higher than could be released from any conceivable process which could take place in any of the constituents of the photographic emulsion.

Hence the only remaining explanation was that the first particle had been converted into another of smaller mass, and the difference of mass had become available as energy. By studying the appearance of the tracks, the masses of both particles could be estimated.

Since then it has become possible to produce these new π mesons in the laboratory, and this has made it more convenient to work with them, and it is also possible to explain the evidence for their behaviour more directly.

This has made it easy to determine accurately the mass of the π meson. It turns out to be about 272 electron masses (as compared with 210 for the μ meson). It has also been found that the mean life of the π meson is only about 10^{-8} seconds, in other words, a hundred times shorter than that of the μ meson. After that time the π meson breaks up into a μ meson and some other invisible particle, without which momentum could not be balanced. Here, only one neutral particle is produced (as compared to the two produced in the decay of the μ meson); it is probably a neutrino, though this has not yet been proved.

The short lifetime is the reason why π mesons had not been noticed before they were discovered by Powell. In its average lifetime, a π meson, travelling even with light velocity would only cover a distance of about twenty feet and, at lower velocity, correspondingly less. This means that it will, in air or in a cloud chamber, usually decay before it has been stopped by friction, and if it decays in flight it is much harder to recognize what is going on. In any event, since each π meson decays to a μ, and the latter lives, on the average, a hundred times longer, one expects μ mesons to be much more abundant.

The photographic plate, which shows tracks of particles moving through solid matter, is therefore ideally suited to the study of such short-lived particles, both because they are brought to rest in a much shorter time and because by observing the plate under a microscope much shorter lengths of track can be studied.

Let us now return to the process by which mesons are produced in nuclear collisions, either in the cosmic radiation or in the laboratory. It is clear, first of all, that this process requires an energy equal at least to the rest energy of a π meson, which is about 140 MeV. However, if a fast proton collides with another stationary proton, its momentum must be conserved and therefore the particles left over after the collision must, between them, have considerable momentum. This means that they must also have kinetic energy, which is to be added to their rest energy. In fact, the least energy which a proton must have to produce a π meson in a collision with a stationary proton is nearly 300 MeV.

The π-mesons are produced abundantly in collisions with protons or neutrons with nuclei above about 300 MeV and in many laboratories there are now machines in operation which can bring particles to the required energies. A similar phenomenon takes place as the primary cosmic rays enter the atmosphere. In their collisions with the nuclei of the air they produce π mesons. These almost at once decay into μ mesons and these in turn into electrons.

Besides these electrically charged mesons, there exist also uncharged, or neutral, ones. This also was predicted from theoretical speculation before it was confirmed experimentally. The neutral π meson is also interesting in being so far the only new particle whose existence was proved by producing it in a machine before it was discovered in cosmic radiation.

The theoretical reasoning in its favour went as follows: We had seen that Yukawa's ideas explained the force between the neutron and the proton by imagining that the proton could, for example, produce "on credit" a meson, which was then absorbed again by the neutron or vice versa. Now if only charged mesons exist, the proton can produce only a positive meson and, to balance charge, it is thereby converted into a neutron. As the positive meson is taken up by the neutron, a proton must result. The net result is that the neutron and the proton have exchanged their charge, and indeed, as we mentioned in Chapter 10, such exchange forces are known to exist. But the same mechanism could never cause any interaction between two like particles, for example, two protons. If a proton produces a positive meson this cannot be taken up by another proton without converting it into a particle with two positive charges, which does not seem to exist. Alternatively, a proton could not produce a negative meson without again being left with two positive units of charge. On the other hand, it is known that the nuclear force between two protons is quite as strong as that between proton and neutron and the easiest way to provide for this is to postulate that not only charged but also neutral mesons exist.

These were indeed found in experiments with fast protons. Their discovery required a subtle piece of detective work, since it turns out that their average lifetime is even shorter than that of the π mesons, in fact, they live only for about 10^{-14} seconds. Even when travelling with light velocity, they therefore cover only about $\frac{1}{1000}$ mm. before they disintegrate. When they do, the result is two photons which, if the neutral π meson was moving slowly, must come out in opposite directions with equal energies, each having about 70 MeV. These photons in turn are detected by means of the electron pairs which they produce in passing through matter. It will be seen therefore that what the observations record is twice removed from the neutral π meson. Nevertheless the experiment first performed by Panofsky and other demonstrates the point convincingly. Later the existence of the same particles has also been demonstrated in the cosmic radiation.

The mass of the neutral π mesons is slightly less than that of the charged ones, and therefore when a charged π meson collides with a nucleus, a neutral π of slightly larger kinetic energy may come out as a result.

Both charged and neutral π mesons have been proved to have no spin. The evidence for this is too technical to be reported here. This result fits in well with the view of Yukawa since it is essential for this that a proton should be able to change into a neutron plus a π meson and vice versa. Since neutron and proton each have a spin of half a unit, this is possible if the π meson has no spin. As far as this argument goes it could still have a spin of a whole number of units, since the balance could always be made up by the motion of the particles in their orbit, but a spin of half a unit would not do.

This result is of some interest in connection with the earlier difficulties in finding a wave description for charged particles without spin. Fortunately, it had been shown long before mesons were discovered, that the fourfold wave equation of Dirac, which required a spin of half a unit, was not the only possible law for waves associated with charged particles in relativity, so that the discovery of charged spinless particles in itself presented no difficulty.

At the present time several laboratories are studying the behaviour of π mesons, in particular in their collision with neutrons and protons. The subject is rapidly passing from the stage of the discovery of new processes to a detailed quantitative survey. One might have hoped that these discoveries, which confirmed in so many ways the ideas put forward by Yukawa to explain nuclear forces, would now have cleared the way for putting Yukawa's ideas into the form of quantitative laws. Much to our disappointment little progress has so far been made in that direction because of new and unexpected difficulties.

The major one of these is due to the fact that the meson field surrounding a proton or neutron is very much stronger than the electromagnetic field surrounding an electron. This is related to the fact that the nuclear force at very short distances is much stronger than electric forces between electrons or protons at corresponding distances, a fact which has already been stressed in Chapter 10. This complicates a detailed description of what goes on. For example, when we discussed, in Chapter 9, the interplay of an electron with the photons which it was capable of producing and reabsorbing, we could be certain that the chance of there being two such photons present simultaneously was, for most purposes, negligible. Similarly, in close collisions of two electrons or an electron and a proton, there would be, in most cases, no energetic photon emitted, only occasionally one, and very rarely two. All this has to do with the fact that the number $\frac{2\pi e^2}{hc}$, which we introduced in Chapter 9, and which measured the interaction of the electron with the surrounding field, is very small, about $\frac{1}{137}$. The corresponding number for a meson is very much larger, probably greater than 1. As a result, we must expect that a proton or neutron will borrow just as readily enough energy to produce two, three or more π mesons as it will produce one. Indeed, in collisions in which there is enough energy available to produce more than one real meson, as in recent experiments at the Brookhaven laboratory, with a machine producing protons at 3,000 MeV, there are more cases in which a collision produces two mesons than one.

As a result, the problem of what consequences one should expect from some particular form of Yukawa-type laws remains unsolved. We do not know whether the next step forward will be some ingenious mathematical discovery which will help us to throw light on these difficult problems, or some new clue derived from experiments which lead us to modify the laws themselves.

BEYOND THE MESONS. A NEW CHAPTER OPENS

The discovery of the μ mesons and the charged and neutral π's in a way agreed with our expectation because the μ meson appeared to be, and the π meson actually was, the confirmation of a bold but very convincing speculation. There was no particular reason to expect that discoveries of new particles would continue. Yet they did continue, and research has already shown up a bewildering variety of new particles which seems to go on growing as the accuracy and ingenuity of the experiments improves.

One of these particles (in fact, probably the first of this group to be discovered) is now called the Λ (lambda) particle, which is neutral, and decays into a proton and a negative π meson. In this decay an energy of 37 MeV is released, in other words the Λ particle has a rest mass equal to that of a proton plus a π meson plus 72 times the electron mass. Its mean life is about 10^{-10} seconds. It occurs in the cosmic radiation. The first example of this type was found in a cloud chamber photograph at sea level, but this is very rare and most of these particles have been seen in cloud chamber photographs on top of high mountains.

Recently, it has been found that such Λ particles can also be produced artificially. They are found when neutrons or π mesons of high energy (produced by the machine in Brookhaven) pass through matter. The process in which they are produced is probably complicated, and there is evidence that, together with this new particle, another new one is produced. This is, at least in some instances, a particle called $\Theta_0{}'$ (theta) which has also been found in cosmic rays. This is another neutral particle, with also a mean life of about 10^{-10} seconds which splits into two π mesons, liberating an energy of just over 200 MeV. In other words, its rest mass is just under 1000 electron masses. A third new particle, the so-called τ (tau), is charged, with a lifetime similar to the others. It decays into three charged π mesons, the energy left over in this process is about 70 MeV.

There are others, for example, the "κ" (kappa), a charged particle which decays into a μ meson and one or two neutral particles, and another the so-called χ (chi), which decays into a charged π meson and a neutral particle.

A good deal of progress is now being made cataloguing these particles and their behaviour, but beyond assigning names to them there is still no clear picture about their connections, nor is there any indication how much further this list will extend.

It has to be borne in mind that lifetimes of 10^{-10} seconds which seem very short on the time scale to which we are used, are still very long times in relation to the frequencies which are associated with such particles. It would therefore not be at all surprising if there existed many more new particles of a much shorter lifetime, which would be harder to detect.

We have met one example of such a short-lived particle, the neutral π meson, which lives only such a short time that the place where it is produced and the place where it decays are distinguishable only by the most refined techniques. Its existence can therefore be proved only by circumstantial evidence, but this was particularly easy in the case of the neutral π meson because, as was mentioned before, there were theoretical reasons to expect such a particle to exist and also some indications of what it would do.

One particle which has not yet been discovered is confidently predicted by theory: that is the negative proton, also sometimes called "anti-proton". To understand the reason for this, we have only to remember the discussion in Chapter 9 of the apparent difficulty over electrons in states of negative energy and the way in which this led to the idea of positive electrons or "anti-electrons". These were the counterpart of the electron in that a positive electron was in some sense the same thing as the absence of an ordinary electron from a state of negative energy. More practically, a positron together with an ordinary electron could disappear, their energy being turned into radiation.

As regards the proton, the situation is quite similar. The laws of proton waves would not make sense and would lead to protons in states of negative energy with quite an absurd behaviour, unless we avoided this trouble in just the same way as it was avoided for electrons, i.e. unless we assumed the existence of anti-protons, or protons with a negative charge, which together with ordinary protons could annihilate, their mass being converted into energy. (It is most likely that this energy would usually come out in the form of mesons.) To create such a proton pair requires an energy of roughly 2,000 MeV. If they are created by the impact of one fast proton on another at rest, as is the normal situation both in laboratory experiments and in collisions of cosmic rays, the incident proton must have rather more energy, since the four particles remaining after the collision (the two protons and the new proton pair) must still be moving to conserve momentum. If one works this out, one finds that the kinetic energy of the proton must be at least six times its rest energy, in other words, about 6,000 MeV. Such energies are not as yet available in the laboratory, but machines are now under construction which will exceed this limit; they should provide an answer to the question whether our speculation about the negative proton is correct.

In cosmic radiation, protons of far greater energies occur and one would have expected that they must, in fact, sometimes produce negative protons. However, these would rarely have been seen in conditions when they are easily identified. For example, since they are produced in extremely violent collisions, they would, in most cases, be moving so fast that they would not be appreciably deflected in any of the magnets that have been used and, hence, they are not easily recognized as negative. One would therefore not see anything unusual in the track of such a particle unless one happened to catch it very close to the end of its range, when it is about to stop, and this is not a frequent occurrence. The fact that no negative proton has yet been seen in cosmic rays is therefore not evidence against its existence.

With these brief remarks we shall have to leave the unfinished story of the manifold new particles. With them we have clearly reached the unexplored frontier of physics. Their relation to the laws of nature, which has been our main theme, is rather negative in that they serve to show up the incompleteness of the laws of nature as we know them to-day. There would seem little doubt that as our understanding of these recent discoveries advances, both by further experimental studies and by realizing the laws which underlie their results, we shall find our present form of the laws of nature not wrong, but incomplete.

Where Do We Stand?

OUR summary of the laws of nature has taken us from the beginning of physics as an exact science to the most recent discoveries, from the law of inertia to collisions of mesons.

It has already been stressed that each new step in this development must not be regarded as disproving what was known before, but as extending and enriching it, so that, in general, the older work retains its value in its proper sphere. It is true that some past work was later shown to be mistaken, based on inaccurate data or faulty reasoning, or on hypotheses which were not confirmed. In our brief account, such work has not been mentioned at all, though it is well to remember that not every idea of even a great physicist stands up to the test, and that many have to be tried before one will prove itself.

Another general experience is that, as our knowledge grows by the solution of some problems, new problems and difficulties arise. The questions with which we were concerned in the later chapters could not have been raised even as questions without the knowledge derived from the results of earlier chapters. As a result, throughout the development of the subject, the magnitude of the unsolved problems has at some times appeared enormous and at others very slight. At some times, new discoveries, which did not fit any recognizable pattern, have followed each other in rapid succession, at others it looked as if we were near to a complete knowledge of the laws of physics. The success of Newtonian mechanics led to an optimistic view that all behaviour of matter could, ultimately, be reduced to the laws of mechanics, and even Maxwell tried to back up his famous laws of the electromagnetic field by mechanical models. Only later, when there had been time to become accustomed to Maxwell's laws and when the ideas of relativity had thrown more light on them, came the recognition that these laws were just as simple in themselves as the mechanical laws and that there was no need to reduce one to the other.

When the mechanical and the electromagnetic laws were clarified, it was thought by some people that the laws of physics were complete. It was a common tendency of text-books, written about the beginning of this century, to present physics as essentially a closed subject, without any reference to unsolved problems. Of course, it was well known that the characteristics of the various substances could not be derived then from any of the known laws, but this was not regarded as part of the task. The specific weight of lead or the electric resistance of copper, the hardness of diamond or the colour of iodine were data which the physicist had to use as his raw material but which he was not trying to explain.

It was realized only later that the laws of physics were applicable to the structure of matter, and that one had to find the basic laws responsible for all this variety. We saw in Chapters 4 and 5 how these problems gradually took shape and, in Chapters 7 and 8, how the formulation of quantum mechanics gave an almost complete account of the structure of matter, with the exception mainly of the interior of the nucleus.

Consequently, there was again a time when the problem of the basic laws seemed simple and a complete solution near achievement. It was realized that the picture was not complete until one could deal with the nucleus, but so little was then known about the behaviour of nuclei that it seemed also that not much was required to explain them. At that time (before the discovery of the neutron) they were believed to contain protons and electrons. Admittedly, one did not know how such a combination of electrons and protons could hold together, but the study of electrons at high speed was giving trouble anyway and one expected the solutions to both problems to be coupled. Only two elementary particles were then known, the proton and the electron, and besides the ratio of their masses, the only other pure number (i.e. the only combination of natural constants which is independent of the units in which we express them) was the fine structure constant $\frac{2\pi e^2}{hc}$. Perhaps not much development would be required before one could explain these two numbers, i.e. explain why all electrons and protons had just the amount of electric charge which they do have and why protons are just about 1840 times heavier than electrons. Perhaps in a few more years, all the basic laws of physics might be known.

I do not want to suggest that this view was seriously held by many physicists at that time and, in fact, most of them realized that deeper knowledge might bring with it new problems. But it was a view which, from the facts known at the time, could reasonably be held. Since then about 25 years have passed and, from having a look of near completeness, basic physics has changed again to a field with more open questions than final answers, as I have tried to bring out in the last two chapters. The list of new particles is steadily growing and with them the list of numbers (such as the relations between their masses, their lifetimes and their interactions), so that there is, at the moment, no shortage of data on which to test any new theory. But, as yet, there is no clue that has helped anyone to propose such a theory. The next step in bringing order into this apparent confusion may be taken to-morrow and it may be delayed for many years. Any speculation about what will come after that can only be a wild guess. It is possible that one day we shall complete our knowledge of the basic laws of physics so that physics will become a closed subject based on known principles, in the sense in which Newtonian mechanics, as applied to objects of everyday life, is a closed subject to-day. But it may well be that this final completion will always remain beyond our reach and that, however far our resources, our ingenuity and our curiosity can carry us along, every new piece of knowledge will reveal beyond it unsuspected new problems.

Even if one day we find our knowledge of the basic laws concerning inanimate matter to be complete, this would not mean that we had "explained" all of inanimate nature. All we should have done is to show that all the complex phenomena of our experience are derived from some simple basic laws. But how to explain the laws themselves? In the days when it appeared that the only basic laws were those of mechanics, one could be under the illusion that these laws were self-evident, though in fact, the readiness of our mind to accept them stems only from being used to them in the behaviour of material objects in day-to-day life.

The first question with which our curiosity launches us on the study of objects around us begins with a "why" and while we go on to trace each fact to a deeper cause, the basic "why" remains, at the end, unanswered. It could not be otherwise, for what we mean by an explanation, by an answer to a "why", is always to demonstrate that the fact in question follows from some accepted law or principle. All explanations of natural phenomena therefore consist in reducing them to some basic laws. To ask for an explanation of these laws would merely mean reducing them to some other laws.

Sometimes people inclined to philosophic generalizations attempt to show that the basic laws, such as those of Newton, are really self-evident from some alleged general principles of which we can be sure even apart from any knowledge of the world around us. This is a rather risky undertaking because the most beautiful derivation of a law of nature from abstract principles must collapse if it turns out that the law itself has to be modified in the light of later discoveries. If it should ever come about that our knowledge of the laws of physics turns out to be complete, then it will, of course, be quite safe, for anyone who wishes, to prove that they could not possibly be any different from what they are.

But to the scientist they will remain what they are now, laws derived from our experience, tested by seeing their predictions come true in new situations, accepted subject to further revision where necessary, as long as they help us to bring order into what we observe.

We may be prejudiced in believing that ultimately the right laws should be simple. We tend to derive an intellectual satisfaction whenever a simple law turns out to fit the facts, though what is simple and what is complicated is, in part, a matter of habits of thought, so that no test of simplicity is absolute.

This book should not end without stressing again the important things that have been left out. No reference has been made to cosmology, i.e. the problem of the structure of the universe as a whole, nor of the related field of cosmogony, the problem of the history and origin of the universe as we see it. There have been points of contact. In looking at the lists of nuclei or isotopes found in nature, one cannot help wondering why some are more abundant than others and the answer to this lies evidently in the process by which, at some earlier time, these nuclei were formed. Or if, as some authors claim, the universe remains in a steady state, then the abundance of various elements or isotopes must depend on the way in which the steady state is maintained. In using the cosmic rays as a source of particles of the highest energy, one cannot help speculating about where they might come from, though this is not essential for a study of what they do.

The main excuse for omitting any account of such problems is lack of space, together with the fact that these subjects are not yet linked with the material we have reviewed in the sense in which relativity and quantum theory are linked. Moreover, it is fair to say that, in these subjects, there is still more speculation and hypothesis than confirmed facts. The many interesting discoveries about the universe have not yet led to laws that can be stated with as much confidence as those which have been described in this book.

A very much wider field that we have not touched is that of life. We have talked of nature as if it contained no plants, no animals and no human beings. The same excuses apply as before and it is probably fair to say of the biological sciences, too, that they have not yet reached the stage at which the simple basic principles begin to become apparent, though it may well be that experts in the biological field would dispute this statement.

It is, however, relevant to speculate about what might turn out to be the relation between the laws of inanimate matter which we have discussed and any laws concerning living matter that might be developed in the future.

As far as is known, living matter seems to conform to the general laws of physics wherever they can be tested. The motion of our limbs seems to follow the laws of mechanics like those of the parts of a machine. The work we do and the heat we produce seem to be balanced by the fuel value of the raw materials in our food. Chemical reactions which can be identified in the living body proceed as they would in a test tube.

But this is not to say that the laws of inanimate physics would be capable of giving a complete account of a living organism. It seems to me probable that the contrary is true and that, in the living being, something manifests itself which, in our description of inanimate nature, is left out of account. These special features may exist only when we are dealing with living matter. Or else it may be that they are always present but are of no importance for the behaviour of inanimate matter and therefore have not, as yet, shown up there. This last possibility would be similar to the way in which the corrections due to relativity exist in principle also for slow objects. A railway train does, in principle, suffer a Lorentz contraction, and there is a slight ambiguity about the timing of events as between a passenger in the train and one standing still, but the differences are so minute that it would be futile to remember them when dealing with actual trains. Similarly, the quantum laws apply, in principle, to a tennis ball, but the corrections, such as diffraction effects, to which they give rise will ordinarily be so minute that it would be unreasonable to take them into account. Similarly, it might be that the new laws which have to be invoked to account for living matter, would also, in principle, modify our description of the railway train, the tennis ball or the atom; but that they would, in these cases, cause only negligible changes and that they would be of importance only in living matter. We cannot, as yet, be sure which view is right.

In any event, our thoughts about the relations between life and physics are bound to be influenced by the changes that have taken place in physics in the last fifty years. Before that, it was a general belief that a complete deterministic description of nature was possible, that we could imagine the position and speed of every atom and of each electron in every atom specified at some moment. Then the laws of mechanics and electricity would allow one to work out the state of affairs at any other time. In such a deterministic picture the existence of living matter would be a complication. If we followed the behaviour of the molecules, atoms and electrons inside a living body in the same manner, would they move exactly as described by the ordinary laws of physics and, if so, could life be just a complicated result of physical laws like a calculating machine or a steam engine? Or, if some molecules of some atoms disobeyed the laws of physics, why did they do so only in the living body? We can learn from quantum mechanics that these questions are meaningless. We can never obtain sufficient knowledge about the motion of even a single atom to forecast its behaviour with certainty in the future, and, as we have seen in Chapter 7, we get involved in contradictions if we postulate that questions have a meaning without specifying the observations by means of which we can put the answers to a test. In living matter this reasoning goes beyond the limitations of the uncertainty principle of quantum mechanics.

The more we want to know about the state of the atoms inside a living organism, the more drastic must be the disturbance which our measurement will cause; and beyond the uncertainties which this will introduce in other physical data, as the uncertainty principle requires, there is the stronger limitation that a living being can put up with only a certain amount of interference before it will cease to be a living being. That this limitation is not by any means academic is illustrated by the fact that the smallest and probably, in some sense, simplest objects of study of the biologist, such as viruses, are made visible only by means of an electron microscope. What we see in such a picture is then not the living virus, since it cannot survive the bombardment with fast electrons to which it must be exposed. What we see are its remains. These, of course, provide important information. But one can see the hopelessness of trying to watch the virus in a microscope of such magnification that one can "see it work".

In this way, while the newer developments in physics are not likely to be of direct help to the biologist they are at least useful in removing misconceptions due to the older mechanistic ideas and widen the possibilities open to the imagination.

It is never possible to predict developments of science before they have happened, but it is not unreasonable to hope, in the near future, for some new insight relating to living matter which, like the discovery of the atom and the development of the quantum theory, would suddenly co-ordinate many facts so far apparently independent and inexplicable.

LIST OF SYMBOLS

This list contains only symbols which recur frequently in the text.

c	velocity of light.
e	electric charge, particularly of electron.
E	electric field strength (in chapter 2). energy.
F	force.
g	acceleration due to gravity.
h	Planck's constant (p. 147).
H	magnetic field strength.
i	electric current.
k	Boltzmann's constant.
l	angular momentum (p. 186).
m	mass. also angular momentum about given axis (p. 184).
n	quantum number (p. 162).
p	momentum.
q,Q	electric charge.
r	distance, radius of curvature.
T	absolute temperature (p. 107)
u,v	velocity.
V	potential energy or electric potential.
x	coordinate.
Z	atomic number (p. 97).
λ	("lambda") wavelength.
ν	("nu") frequency.

INDEX

(Page numbers in bold type indicate the main reference, or the explanation of a term)

Absolute temperature, 108
absorption of light, 156
abundance of isotopes, 253
acceleration, 20
action, 21
— at a distance, 50
aerial, 67
alchemists, 253
alkali, 97
alpha particle, **95**, 230, 245
alpha rays, **95**, 226
Anderson, 262
angular momentum, **35**, 183, 246
anti-proton, 271
Aston, 239
atom, 88
atomic, bomb, 251
— energy, 226
— hypothesis, 88
— number, 97
— power, 251
— weight, **88**, 238
attraction, 24, 41, 240
average, 109

Barrier, 205
battery, 30, **52**, 92
beats, 83
Becquerel, 95
beryllium, 181, 192
beta-decay, **246**, 262
beta rays, **95**, 245
billiard balls, 31
binding energy, 240
black-body radiation, **118**, 147
Blackett, 233, 262
Bohr, 9, 163, 169
Boltzmann's constant, 109
bond, 198
boron, 181, 192
boundary conditions, 39
break-up, 227
Brookhaven, 269
Brownian motion, 111
BTU, 102

Calculus, 20
caloris, 102
cancellation, 166
carbon, 30, 181, 192, 198
cascade, 259
cathode rays, 94
Cavendish, 25
causality, 171
centre of mass, 26
centrifugal force, **23**, 38
Chadwick, 237
Champion, 139
charge, 41
— independence, 243
chemical energy, 30
chi particle, 270
chlorine, 238
circulation, **30**, 43, 48, 63
— magnetic, 54
classical physics, **121**, 173
cloud chamber, 233
coal, 30
Cockcroft and Walton, 235
coincidence, 256
cold emission, 208
collision, 31
— of molecules, 106
— in relativity, 134
— with nucleus, 233
commutator, 56
composition of velocities, 133
compound, 87
Compton effect, **150**, 168
condenser, 48
Condon and Gurney, 230
conductor, **44**, 101
conservation of angular momentum, 36
— of energy, 28, 30, 134, 246
— of mass, 88
— of momentum, **26**, 134
conservative force, **28**, 43
contact force, 25, 31
cosmic rays, 255
cosmology, 276
Coulomb's law, **41**, 48, 218, 241
counter, **235**, 256
crystal, 115
Curie, 95
current, 52
curvature, **22**, 38, 159
cyclotron, 236

Davison and Germer, 164
de Broglie, **153**, 163
Debye, 204
decay product, 228
degrees of freedom, **39**, 119
deuterium, 239
deuteron, **239**, 242
diamond, 200
dielectric constant, 69, **100**
differential calculus, 20
— equations, 20, 158
diffraction, **75**, 165
— of electrons, 152
— of X-rays, 90
dimensions, 39
Dirac, 201, 213
discharge tube, 179
disorder, 105
displacement current, 63
— of stars, 144
divergence-free, **48**, 54, 63
Dulong-Petit, **116**, 203

Eclipse, 144
Eddington, 110
education, 8
Einstein, 130, 149, 204
elastic collision, 32
— energy, 32
— forces, 25
electric charge, 41
— energy, 30
— field strength, 42
— flux, 63
— heater, 30, 52
— potential, 43
electricity, 41
electro-positive, 195
— negative, 195
electrolysis, 92
electromagnetic unit, 55
— mass, 221
— wave, 64
electromotive force, 59
electron, 93
— microscope, 278
electroscope, **45**, 255
electrostatics, 43
elements, 87
emission of light, **156**, 224
emulsion, 258
energy, 28
— levels, 155
entropy, 104
equations of motion, 38
ether, 51
— wind, 124
evaporation, 117
exchange forces, 244
exclusion principle, **182**, 201, 214, 242
explanation, 275

Families of radioactive elements, 228
falling object, 23
Faraday, 92
— cage, 45
— law of induction, 57
Fermi, 201
field energy, 68
— strength, 42
fine structure, 222
— — constant, 224, 269, 274
first law of thermodynamics, 105
fission, 249
fluorine, 181, 192
flux, 57
force, 19
Foucault pendulum, 124
Fraunhofer lines, 156
free energy, 104
frequency, **58**, 147, 156
Fresnel, 72
friction, 20, **29**, 52
fringes, 72
fundamental vibration, 154

Galileo, 20
galvanometer, 55
gamma rays, 80, **95**, 150, 245
Gauss, 47
Gay-Lussac, 107
Geiger counter, 235
— Müller tube, 235
general relativity, 141
generator, 57
geometrical optics, 80, **86**, 173
Goudsmit, 191
grating, 75
gravitation, **24**, 41, 143
gravity, **23**, 28, 142
ground state, 162

Halogen, 97
head-on collision, 33
heat, 102
— conduction, 103
— energy, 30
heavy hydrogen, 239
Heisenberg, 9, 163, 169
helium, 95, **178**, 191
—liquid, 205
Hertz, 67
Hess, 256
history of science, 8
— of the universe, 253

Hogben, 9
hole, 216
Huygens, 72
hydrogen atom, 97, **161**, 191, 222
— molecule, 196

Indivisible atom, 228
induction, 57
inelastic collision, **32**, 137
inertia, 21
infra-red, 80
initial conditions, 38
intensity, 42
interference, 71, 166
inverse-square law, **24**, 41
ion, 92, **94**, 179, 195
ionic compound, 196
irreversible, 103
irrotational, **48**, 53, 63
isotopes, 239
— of uranium, 250

Joliot-Curie, 248

K shell, 193
kappa particle, 270
Kepler, 24
kinetic energy, **28**, 137
— theory, 105

L shell, 193
Lamb, 222
lambda particle, 270
lamp, 52
lattice, 115
Laue, 90
Lawrence, 236
length in relativity, 130
lever, 34, 37
life, 15, 276
line of action, 35
— of force, 45
— spectrum, 162
linear accelerator, 235
liquid, 116, 204
lithium, 180, 191
local laws, 51
longitudinal part, 35
— wave, 73
Lorentz, 126, 219
— contraction, **126**, 263

M shell, 193
Mach, 90
magnet, 54
magnetic energy, 60 **68**
— field of current, 52
— pole, **42**, 54
magnetism, 42
magnetostatics, 43
mass, 21
— defect, 240
— spectrograph, 239
Maxwell, 61
mechanical energy, 30
— mass, 221
mechanics, 19
melting, 116
Mendeleyev, 97
meson, 261
π — 265
μ — 265
metal, 200
methane, 198
MeV = million electron volts, 228
Michelson and Morley, 125
micro-crystalline, 115
microwaves, 222
Millikan, 94
mixture, 87
molecule, 88, 195
moment, 34
momentum, **25**, 135
monochromatic light, 71
motion, 19
motor, 30, **56**
mu-meson, 265

Negative charge, 41
— energy, 211, 271
— proton, 271
neon, 181, 193
neutral π meson, 267
neutrino, 247, 262, 266
neutron, 237, 248
— from fission, 250
Newton's second law, 20
— third law, 21
— law of gravitation, 24, 43
— theory of light, 70
nitrogen, 181
noble gas, 97, 181, 198
node, 65
north pole, 42
nuclear forces, 240
nucleus, **97**, 225

Oppenheimer, 9
orbit of the earth, 146
— of electron, 98, 170, **176**
— of planet, 24, 98

organic chemistry, 198
origin of matter, 253
oxygen, 30, 181

Panofsky, 268
pair annihilation, 217
— creation, 217, 259
parallelogram of forces, 22
Pauli, 9, 182, 247
penetration of barrier, **206**, 230
periodic system, 97, 193
phase, 77
philosophy of science, 8
photo-electric effect, **149**, 235
photographic plate, **258**, 265
photon, 152, 218
pi meson, 265
Planck, 147
planet, 24, 37
polarization, **67**, 77
polaroid, 77
pole, **42**, 54
position 19
positive charge, 41
— electron, **216**, 271
positron, 217
positronium, 217
potential, 43
— barrier, 205, 230, 249
— energy, 28, 37
Powell, 265
pressure, 25, **106**
primary cosmic radiation, 260
principle of equivalence, 141
probability, 110, 165, 172
proportions, constant, 88
proton, **98**, 161, 232, 238

Quantity of heat, 102
quantum, 147
— number, 178
— of action, 147
— theory, 16, 147

Radiant heat, 79, **118**
radio aerial, 67
radioactivity, 95, 226
radium, 95, 226
— emanation, 227
rainbow, 70
rate of change, 19
reaction, **21**, 26, 134
reactor, 251
red-shift, 144
refraction, 71
relative motion, 32
relativity, 16, **121**
repulsion, 41
resistance, 20
— electric, 52
rest energy, **138**, 263
— mass, 136
resultant force, 22
retardation, 218
rigid body, 34, 39
— — in relativity, 219
ripple, 72
Rømer, 78
rotation of molecule, 112
Rutherford, 95, 226, 231
Rydberg constant, 162

Scattering of neutrons, 242
Schrödinger, 163
— equation, 158
Schwinger, 223
scintillation screen, 233, 235
second law of thermodynamics, 105
self-energy, 219
— field, 223
— induction, 60
shell model, 244
— in atom, 193
short-range force, 241
slowing-down, 232
sodium, 181, 193
— chloride, 195, 199
— light, 71
Soddy, 238
soft component, 259
south pole, 42
space in relativity, 130, 143
special relativity, 141
specific heat, 103,
— — of a gas, 112
— — of metals, 202
— — of solids, 116, 204
spectrograph, 77
spectrum, 78
speed, 19
— of light, 78
spin, **182**, 213, 246, 264, 268
spinning top, 37
splitting of nucleus, 231
spontaneous fission, 249
starting conditions, 38
statistical description, 110
— mechanics, 105
steam engine, 29
Stern-Gerlach, 185, 189
sun, source of heat, 251
superconductivity, 203

Tau particle, 270
temperature, 102
thermonuclear process, 252
— bomb, 253
theta particle, 270
Thomson, G. P., 164
Thomson, J. J., 93
time, 19
— dilation, **133**, 145, 263
— in relativity, 130
transformer, 58
transverse part, 35
— wave, 67, **73**
triangle rule, 22, 33

Uhlenbeck, 191
ultra-violet, 80
uncertainty principle, **169** 174, 185, 206
— statistical, 110
uniform motion, 19
uranium, 226, 249

Valency, 198
Van de Graaff, 235
velocity, 19
— of light, 66
— of wave, 66
vector, 22
vibration of molecule, 113
virtual state, **209**, 218, 265
virus, 278
voltage, 59

Wave, 64, 72
— amplitude, 167
— crest, 72
— function, 167
— mechanics, 159
— number, 83
— packet, **83**, 170
— train, 83
— trough, 72
wavelength, 65, **72**, 152, 153
— variable, 158
weight, 24
— of atom, 91
Wilson, 233
work, **28**, 43

X-rays, **80**, 150

Yukawa, 264

Zeeman effect, 191
zero-point motion, 156, 205